Technische Universität München

Fakultät für Mathematik

Quantum Learning Theory

Matthias C. Caro

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Felix Krahmer
Prüfer der Dissertation:	
	1. Prof. Dr. Michael M. Wolf
	2. Prof. Dr. Jens Eisert
	3. Assist.-Prof. Dr. Richard Kueng

Die Dissertation wurde am 19.01.2022 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 19.07.2022 angenommen.

Für meine Eltern.

Zusammenfassung

Diese Dissertation befasst sich mit Fragen der Quantenlerntheorie, an der Schnittstelle von Quanteninformationstheorie und maschinellem Lernen. Wir beweisen Garantien darüber, wie variationelle Quantenlernmodelle von Trainingsdaten auf unbeobachtete Daten verallgemeinern. Und wir untersuchen Lernprobleme, in denen von Quantendaten gelernt wird. Schließlich diskutieren wir Fragen zur Markovianität von Quantenevolutionen und Aspekte von Unentscheidbarkeit in der Lerntheorie.

Abstract

This dissertation treats questions from quantum learning theory, at the intersection point of quantum information theory and machine learning. We prove guarantees on how variational quantum machine learning models generalize from training data to unseen data. And we explore the complexity of tasks of learning an unknown map from quantum data. Finally, we discuss questions related to Markovianity in quantum evolutions and aspects of undecidability in learning theory.

“ *We can only see a short distance ahead, but we can see plenty there that needs to be done.* ”

——— Alan Turing

Acknowledgments

Among the many people who have supported me on my way towards completing this dissertation, first and foremost I want to thank my advisor Prof. Dr. Michael M. Wolf. Thank you for your encouragement, your advice and guidance, your knowledge and wisdom, and for your patience with me in our discussions!

Furthermore, I thank Prof. Dr. Jens Eisert, and Assist.-Prof. Dr. Richard Küng for being part of my examining committee and for taking the time to read my thesis. Special thanks go to Prof. Dr. Felix Krahmer for acting as chair of my examining committee.

One of the most enriching aspects of my doctoral studies was the M5 quantum information group. Prof. Dr. Michael M. Wolf and Prof. Dr. Robert König have created a work environment that I always enjoyed being a part of, Silvia Schulz made sure that everything actually worked even while we mathematicians had our heads in the clouds, and all the different doctoral candidates, postdocs, and visitors – too many to name them all here – contributed to a vibrant atmosphere and fun collaborations. I am honored to have been a part of this group and I hope to often be one of its visitors in the future!

Over the past years, I have also had the pleasure of working together with many amazing researchers. Warm thanks to all of my collaborators, both past and present! To me, research is most fun as a team sport, and I have had – and still have – the good fortune of amazing teammates. Among my collaborators, I am particularly grateful to Benedikt R. Graswald and Markus Hasenöhrl for proofreading this thesis, thereby helping me express my thoughts in a readable manner, and for many musings on the trials and tribulations of a mathematics doctorate.

As a doctoral candidate, I had opportunities to visit the Theoretical Quantum Optics Group at the University of Siegen, the Quantum Information Theory Group at the Università degli Studi di Pavia, the Institute for Integrated Circuits at the JKU Linz, the QMATH Centre at the Københavns Universitet, and the Laboratoire de Physique Théorique de Toulouse. Huge thanks to Gael Sentís, Otfried Gühne, Alessandro Bisio, Paolo Perinotti, Robert Wille, Richard Küng, Andreas Bluhm, Albert Werner, and Ion Nechita for their hospitality and the exchange of ideas.

I gratefully acknowledge the funding that I received for my doctoral studies from the Elite Network of Bavaria through the TopMath Program and from the Studienstiftung des deutschen Volkes. And I thank Agnieszka Baumgärtel, Prof. Dr. Martin Brokate, Prof. Dr. Marco Cicalese, Dr. Carl-Friedrich Kreiner, and Dr. Katja Kröss. Without them, TopMath would not be anywhere near as good a program as it is today. Moreover, I thank Dr. Frank Hofmaier for acting as my mentor for the time as doctoral candidate.

Throughout my doctorate, I could always count on the support of my family and friends, and I am immensely grateful for that. In particular, Robert Caro, my grandfather, has always encouraged me and had many life lessons to share with me. But the one who supported me more than anyone else, proofread this thesis, patiently listened to all my problems, cheered me up when I was feeling low, pushed me when I had trouble motivating myself, and made sure I did not forget to celebrate the small victories, was you, Helene. Thank you, Helene Lösl, for your love and support, I could not have done this without you.

List of contributed articles

Core articles as principal author

I) Matthias C. Caro and Ishaun Datta.
Pseudo-dimension of quantum circuits.
Quantum Mach. Intell. 2, 14 (2020). https://doi.org/10.1007/s42484-020-00027-5.
(see also article [1] in the bibliography)

II) Matthias C. Caro and Benedikt R. Graswald.
Necessary criteria for Markovian divisibility of linear maps.
Journal of Mathematical Physics 62, 042203 (2021). https://doi.org/10.1063/5.0031760.
(see also article [2] in the bibliography)

III) Matthias C. Caro.
Binary classification with classical instances and quantum labels.
Quantum Mach. Intell. 3, 18 (2021). https://doi.org/10.1007/s42484-021-00043-z.
(see also article [3] in the bibliography)

IV) Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke.
Encoding-dependent generalization bounds for parametrized quantum circuits.
Quantum 5, 582 (2021). https://doi.org/10.22331/q-2021-11-17-582.
(see also article [4] in the bibliography)

Further articles as principal author

V) Matthias C. Caro.
Quantum learning Boolean linear functions w.r.t. product distributions.
Quantum Inf Process 19, 172 (2020). https://doi.org/10.1007/s11128-020-02661-1.
(see also article [5] in the bibliography)

VI) Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles.
Generalization in quantum machine learning from few training data.
Nature Communications 13, 4919 (2022). https://doi.org/10.1038/s41467-022-32550-3.
(see also article [6] in the bibliography)

Further preprints and articles as principal author under review

VII) Matthias C. Caro.
Undecidability of Learnability.
arXiv preprint arXiv:2106.01382.
(see also article [7] in the bibliography)

Articles as co-author

VIII) Markus Hasenöhrl and Matthias C. Caro.
Quantum and classical dynamical semigroups of superchannels and semicausal channels
Journal of Mathematical Physics 63, 072204 (2022). https://doi.org/10.1063/5.0070635
(see also article [8] in the bibliography)

I, Matthias C. Caro, am the principal author of the Core Articles I, II, III, and IV, of the Articles V and VI, and of the Article VII, which is at the moment of the submission of this dissertation an arXiv preprint. I am a coauthor of Article VIII.

Contents

1 Introduction **1**

1.1 Outline . . . 1

1.2 Summary of Results . . . 2

2 Mathematical Ingredients of Quantum Information Theory **9**

2.1 Quantum States and Measurements . . . 10

2.2 Completely Positive Maps and Quantum Channels . . . 12

2.3 Superchannels, Semicausality, and Semilocalizability . . . 16

3 Mathematical Ingredients of Statistical Learning Theory **19**

3.1 Probably Approximately Correct Learning . . . 19

3.2 Complexity Measures and Generalization Bounds . . . 23

3.3 Relations Between Complexity Measures . . . 29

3.4 Undecidable Problems in Classical Learning Theory . . . 31

4 Variational Quantum Machine Learning **35**

4.1 Parametrized Quantum Circuits for Machine Learning . . . 35

4.2 Generalization Guarantees for Variational Quantum Machine Learning . . . 40

4.2.1 Generalization Guarantees Based on the Trainable Part . . . 40

4.2.2 Generalization Guarantees Based on the Data-Encoding . . . 44

5 Learning from Quantum Data **47**

5.1 Learning Classical Concepts from Quantum Examples . . . 47

5.2 Learning State Preparation Procedures from Quantum Data . . . 52

6 Quantum (Non-)Markovianity **57**

6.1 Quantum Dynamical Semigroups . . . 58

6.2 Infinitesimal Markovian Divisible Quantum Channels . . . 60

6.3 Quantum Dynamical Semigroups of Quantum Superchannels . . . 62

Bibliography **65**

Appendices

A Core Articles **79**
A.1 Pseudo-dimension of quantum circuits 79
A.2 Necessary criteria for Markovian divisibility of linear maps 98
A.3 Binary classification with classical instances and quantum labels 124
A.4 Encoding-dependent generalization bounds for parametrized quantum circuits . . 153

B Further articles as principal author **203**
B.1 Quantum learning Boolean linear functions w.r.t. product distributions 203
B.2 Generalization in quantum machine learning from few training data 249

C Further preprints and articles as principal author under review **297**
C.1 Undecidability of Learnability 297

D Articles as co-author **331**
D.1 Quantum and classical dynamical semigroups of superchannels and semicausal channels 331

Chapter 1

Introduction

Both quantum information and machine learning open up new perspectives on computation, the former by allowing for algorithms exploiting features of quantum mechanics, the latter by introducing learning algorithms as a new paradigm of meta-algorithms. Quantum learning theory aims to understand the potential as well as the limitations of combining these two perspectives. Exactly that is the guiding theme of this thesis.

1.1 Outline

In this section, we explain the quantum information-theoretic and learning-theoretic questions that we explore in this thesis. We also give an overview over the different chapters.
Understanding training data requirements is vital for any machine learning problem, both classical and quantum. In machine learning theory, it has long been recognized that statistical features of learning problems are crucial in determining how much data is necessary and/or sufficient for solving them. In particular, statistical learning theory has considered the question of when training data is representative of the true data distribution through the lens of generalization. For this framework, the following question is central: When is the average performance of a machine learning model on the available data a reliable proxy for its performance on previously unseen data? This question, asked for variational quantum machine learning models based on parametrized quantum circuits, is also one of the focus points of this thesis. We answer it by proving upper bounds on the training data size sufficient for good generalization in terms of different architectural properties of the parametrized quantum circuits used to define a machine learning model. Beyond variational quantum machine learning, we also investigate the training data requirements in two concrete tasks of learning from quantum data. Moreover, we discuss generalization in both classical and quantum machine learning from the perspective of computational and logical undecidability.
While we have become used to the usefulness of data in our modern world, it is not automatic that information available in the present allows us to make accurate predictions about the future when considering physical processes. Namely, in quantum physics, not all processes are Markovian. Here, we call a process Markovian if, given the present, the future is independent of the past. Therefore, we also explore quantum (Non-)Markovianity in this thesis. On the one

hand, we consider a mathematical formalization of quantum Markovianity as a certain divisibility property, called infinitesimal divisibility, and establish necessary criteria for a linear map to satisfy this property. On the other hand, we study Markovianity in a type of higher-order quantum evolutions by characterizing the generators that give rise to continuous one-parameter semigroups of quantum superchannels.

The remainder of this thesis is structured as follows. In Section 1.2, the rest of this chapter, we summarize the contributed articles appearing in this thesis. Chapter 2 gives an introduction into selected topics of quantum information theory. Chapter 3 presents basic notions from classical learning theory and discusses aspects of undecidability in learning theory. In Chapters 4 and 5, we present two topics in the theory of quantum machine learning, at the intersection of quantum information and learning theory, with a focus on theoretical guarantees for the training data requirements. Namely, Chapter 4 revolves around variational quantum machine learning and Chapter 5 investigates different scenarios of learning from quantum data. In Chapter 6, we review the notion of quantum Markovianity arising from the study of continuous one-parameter semigroups, discuss a relaxation in terms of divisibility properties, and extend the framework of continuous one-parameter semigroups from quantum channels to quantum superchannels.

Finally, we include all the contributed articles. Before each article, we summarize the main contributions of the respective work and describe the individual contributions of the author of this thesis. In addition, preceding each article, we include the respective permission to use it in this thesis.

1.2 Summary of Results

The contributed articles deal with different questions in quantum information theory, classical learning theory, and quantum learning theory. Core Article I investigates the pseudo-dimension, a complexity measure from classical learning theory, for function classes describing the statistics of measurements performed at the output of a quantum circuit with variable gates. In a similar spirit, Core Article IV and Article VI bound complexity measures for variational quantum machine learning models based on parametrized quantum circuits, thereby deriving generalization guarantees. Here, Core Article IV emphasizes the effect of the encoding gates in the circuit, Article VI focuses on the influence of the trainable gates on generalization. Both Core Article III and Article V deal with tasks of learning from quantum data. In Core Article III, the goal is to learn an unknown quantum state preparation procedure from classical-quantum training examples. In contrast, Article V revolves around learning classical linear functions from non-uniform quantum superposition examples. Article VII addresses a question at the intersection point of classical learning theory, logic, and computability, namely the (un-)decidability of learnability. Finally, Core Article II and Article VIII investigate questions related to quantum (Non-)Markovianity, the former in terms of a divisibility property for quantum channels, the latter by studying continuous one-parameter semigroups of superchannels. Note: The author of this thesis does not claim to be the principal author of the Article VIII.

Core articles as principal author

- *Article I [1]: Pseudo-dimension of quantum circuits*

  In this work, we propose the pseudo-dimension, a combinatorial complexity measure from classical learning theory, as a tool to quantify the expressivity of 2-local quantum circuits on qudits. Here, to a quantum circuit with variable gates we associate a class of $[0, 1]$-valued functions that describe the possible outcome distributions of measurements performed at the output of the circuit, upon input of the $|0\rangle$ qudit state. The main result of this first core article are upper bounds on the pseudo-dimension of this function class. Moreover, we extend the result to more general scenarios of variable circuit circuit architecture, variable input states, and circuits consisting of general quantum operations beyond unitaries. Crucially, our pseudo-dimension bounds scale polynomially in the size and depth of the circuit, as well as the local dimension. In particular, for efficiently implementable quantum circuits, which have depth polynomial in the number of qudits, this complexity bound also scales polynomially in the number of qudits. We demonstrate two applications of our pseudo-dimension bounds. First, we show that polynomial-size training data suffices to probably approximately correctly learn 2-local quantum circuits of known polynomial size and depth. Second, we exhibit an explicit finite class of quantum states at least one of which has exponential gate complexity of state preparation.

  Our proofs rely on understanding the different $[0, 1]$-valued functions implemented by a quantum circuit as arising by fixing some of the variables in a multivariate polynomial. This polynomial is determined by the quantum circuit and its degree can be upper-bounded using the size of the circuit. With this reinterpretation of the function class, bounding its pseudo-dimension becomes a task of upper-bounding the number of consistent sign assignments to a family of polynomials, whose degree we can bound. For this, we employ a result due to [9].

- *Article II [2]: Necessary criteria for Markovian divisibility of linear maps*

  The focus of this second core article is a divisibility notion related to (Non-)Markovianity of quantum evolutions. Ref. [10] introduced the notion of infinitesimal (Markovian) divisibility for quantum channels in. Generalizing this approach, we define Markovian divisibility and infinitesimal Markovian divisibility for general linear maps and sets of generators. We propose a general proof strategy for deriving necessary criteria for (infinitesimal) Markovian divisibility. More precisely, we show how to prove singular value inequalities for (infinitesimal) Markovian divisible maps, given certain spectral properties of the admissible generators. Our proofs are based on Trotterization and on majorization inequalities from matrix analysis.

  Following this strategy for the case of quantum channels and Lindblad generators, we prove the first non-trivial necessary criteria for infinitesimal divisibility of quantum channels that hold in any finite dimension. The main technical ingredient for this proof is a careful analysis of eigenvalues of real parts of Lindblad generators. Through concrete examples,

we show our criteria to be almost optimal. Moreover, we demonstrate that no analogous criteria can hold in general in the classical counterpart of the quantum setting.

- *Article III [3]: Binary Classification with Classical Instances and Quantum Labels*

  In this third core article, we propose a toy problem for learning quantum state preparation procedures from classical-quantum data and characterize its optimal sample complexity. Concretely, we introduce a variant of binary classification in which the labels are quantum states. Importantly, we assume that a learner has access to the quantum labels in the training data only via actual quantum copies of the respective label state, not in terms of a full classical description.

  We show upper bounds on the sufficient training data size for this quantum learning problem in terms of the complexity of the hypothesis class used by the learner. In proving these upper bounds, we use local Holevo-Helstrom measurements to reduce the learning problem to a task of learning from classical data with noisy labels. Assuming pure quantum label states, we complement the sample complexity upper bounds with essentially matching lower bounds. The proofs of these lower bounds rely on a reduction to a problem of quantum state discrimination. We lower bound the sample complexity of the latter via an information-theoretic argument.

- *Article IV [4]: Encoding-dependent generalization bounds for parametrized quantum circuits*

  Variational quantum machine learning is the subarea of quantum machine learning in which parametrized quantum circuits (PQCs), trainable via classical optimization procedures, serve as machine learning models. Applying such models to problems of learning from quantum data requires a way of encoding the classical data such that it can be processed by the quantum circuit. In this fourth core article, we investigate the effects of the quantum data-encoding strategy on the training data requirements for good generalization in PQC-based machine learning models. For several commonly used encoding strategies, we show that training data of size polynomial in the number of encoding gates suffices to guarantee good generalization. Moreover, we demonstrate how to establish generalization bounds for a PQC-based machine learning model by solving a purely combinatorial question about the spectra of the encoding Hamiltonians.

  For our proofs, we use a representation of the functions implemented by a PQC in terms of generalized trigonometric polynomials (GTPs), arising from successive diagonalizations of the Hamiltonians used for encoding the classical data. Crucially, we show that the accessible frequency spectrum in the GTP representation is determined by the data-encoding strategy. This allows us to derive explicitly encoding-dependent generalization guarantees for PQC-based quantum machine learning from generalization guarantees for classes of GTPs that explicitly depend on the accessible frequency spectrum. We demonstrate two proof strategies for establishing the latter. First, we reinterpret GTPs as functions implemented by simple neural networks, for which Rademacher complexity bounds are known. Second, we regard the number of accessible frequencies as a bound on the effective dimension of a class of GTPs and prove covering number bounds on this basis.

Further articles as principal author

- *Article V [5]: Quantum learning Boolean linear functions w.r.t. product distributions*

  In this article, we investigate a problem of learning Boolean linear functions from quantum superposition examples. Prior work [11, 12] had demonstrated that quantum Fourier sampling, in particular the Bernstein-Vazirani algorithm [13], can serve as a basis for a quantum algorithm that efficiently learns linear functions from superposition examples, even in the presence of noise. However, these results required the superposition weights to be uniform. We quantitatively investigate how a bias away from uniformity in the superposition weights influences their usefulness for learning linear functions. More precisely, we show the following results: For small bias, a constant number of quantum examples suffices for exactly learning an unknown Boolean linear function on n bits. For any arbitrary (except full) bias, a number of quantum examples scaling logarithmically in n is sufficient. And for large bias, the quantum sample complexity cannot scale better than logarithmically in n. As classical learners require linearly-in-n many classical examples to exactly learn an unknown Boolean linear function, this shows that even for biased product distributions, quantum learners can outperform classical learners in both sample and computational complexity.

  In proving the sample complexity upper bounds, we build on a biased version of the quantum Fourier transform and a corresponding biased version of quantum Fourier sampling [14]. With this, we define a biased version of the Bernstein-Vazirani algorithm and, through amplification, use it as subroutine for a quantum learning algorithm. To prove the complementary sample complexity lower bounds, we relate the learning task to a problem of identifying an unknown quantum state from a known ensemble. For the latter, we obtain sample complexity lower bounds by bounding the success probability of the so-called "pretty good measurement", through studying the Gram matrix of the ensemble.

- *Article VI [6]: Generalization in quantum machine learning from few training data*

  This article is a continuation of the line of research initiated in Core Article I [1] and at the same time complementary to Core Article IV [4]. Namely, here we investigate the generalization behaviour of variational quantum machine learning models based on parametrized quantum circuits, with a focus on the trainable part of the circuit. We prove an upper bound of the generalization error in variational quantum machine learning in terms of the number of trainable local quantum channels in the quantum circuit. In particular, our result implies that efficiently implementable models, which have polynomially many gates, can be learned from polynomial-size training data. Moreover, we extend our bounds to models with shared parameters between gates, variable circuit architecture, and take the optimization process into account. This makes the bounds flexibly applicable, as we demonstrate with theoretical guarantees for quantum phase recognition and unitary compiling, both confirmed by numerical experiments.

  Our main technical contribution, which is at the core of our proofs, are covering number bounds for the class of quantum channels that a quantum machine learning model can

implement. To obtain these bounds, we combine compactness of the set of local quantum channels with the circuit structure and subaddivity of the distance induced by the diamond norm. We then use Dudley's covering number integral to relate Rademacher complexities and covering numbers, and finally obtain generalization bounds from the Rademacher complexity bounds.

Further preprints and articles as principal author under review

- *Article VII [7]: Undecidability of Learnability*

  This article explores the following fundamental questions in classical learning theory: If learning is possible, can we prove that this is indeed the case? Is there an algorithmic procedure for deciding for any given scenario whether learning is possible? We prove that the answer is negative for both of these questions: For different learning scenarios, learnability is in general undecidable, both in the sense of independence of the axioms in a formal system and in the sense of uncomputability. More precisely, we show this for the models of probably approximately correct binary classification, uniform and universal online learning, and exact learning through teacher-learner interactions.

  In all of these models, learnability is determined by combinatorial properties of the model class. For example, under suitable measurability assumptions, learnability for probably approximately correct binary classification is equivalent to the model class having finite VC-dimension. Similar characterizations of learnability are known for online learning in terms of Littlestone trees, and for teaching problems in terms of the teaching dimension. In our proofs, we show that deciding whether a class satisfies any of these combinatorial properties is in general not possible. We derive this from two fundamental undecidability results, namely Gödel's second incompleteness theorem and the undecidability of the halting problem.

Articles as co-author

- *Article VIII [8]: Quantum and classical dynamical semigroups of superchannels and semi-causal channels*

  The full characterization of the generators of continuous one-parameter semigroups of quantum channels [15, 16], now usually called GKLS- or Lindblad generators, is fundamental for studying (Non-)Markovianity in evolutions of quantum systems. In contrast, no corresponding characterization for the generators giving rise to semigroups of quantum superchannels was known. Here, superchannels are linear maps that map admissible quantum evolutions to admissible quantum evolutions. In this article, we close this gap in the literature by fully characterizing such generators, both via an efficiently checkable criterion, based on a semidefinite program, and via a normal form. These can serve as a basis for both analytical and numerical investigations of (Non-)Markovianity in higher-order quantum theory. In addition to the quantum case, we also resolve the analogous question in the classical case.

As the first step towards proving our normal form, we use a correspondence between superchannels and certain semicausal completely positive maps. Here, semicausal maps on a bipartite system allow information flow between the subsystems only in one direction. Thus, it suffices to characterize the generators of semicausal channels. In the quantum case, we achieve the latter through a technique based on Haar integration, which allows to transfer the semicausality assumption from a whole Lindblad generator to its completely positive part. Now, the known equivalence between semicausal and semilocalizable quantum channels, which we extend to infinite dimensions, leads us to a constructive expression for the admissible generators.

Chapter 2

Mathematical Ingredients of Quantum Information Theory

In this chapter, we introduce the mathematical framework for finite-dimensional quantum information theory. The material presented here can be found in textbooks such as [17–19] or lecture notes such as [20–22]. Section 2.1 discusses the density matrix formalism for describing the state of a (potentially composite) quantum system and the corresponding description of measurements in quantum theory. In Section 2.2, we present the channel formalism for transformations of quantum systems, before moving on to the superchannel formalism for transformations of transformations.

In preparation for the remainder of this chapter and this thesis, let us introduce some notation. Throughout this chapter, we work with finite-dimensional Hilbert spaces $\mathbb{C}^d$, with dimension $d \in \mathbb{N}_{\geq 1} \coloneqq \{1, 2, \ldots\}$. We denote the set of bounded linear operators between finite-dimensional Hilbert spaces $\mathbb{C}^{d_A}$ and $\mathbb{C}^{d_B}$ by $\mathcal{B}(\mathbb{C}^{d_A}; \mathbb{C}^{d_B})$. If $d_A = d_B = d$, we write $\mathcal{B}(\mathbb{C}^d) = \mathcal{B}(\mathbb{C}^{d_A}; \mathbb{C}^{d_B})$. We will also use $\mathcal{M}_{d_B, d_A}$ to denote $d_B \times d_A$ matrices with complex entries, and write $\mathcal{M}_d$ if $d_A = d_B = d$. For $X \in \mathcal{B}(\mathbb{C}^d)$, $\mathrm{tr}[X]$ denotes the trace of (a matrix representing) X. Here and throughout, we sometimes implicitly identify $\mathcal{B}(\mathbb{C}^{d_A}; \mathbb{C}^{d_B})$ and $\mathcal{M}_{d_B, d_A}$. Using the trace, we can equip $\mathcal{B}(\mathbb{C}^{d_A}; \mathbb{C}^{d_B})$ with the Hilbert-Schmidt inner product $\langle X, Y \rangle_{\mathrm{HS}} \coloneqq \mathrm{tr}[X^\dagger Y]$. Here, $X^\dagger$ denotes the conjugate transpose of X. Note that, as is common in the mathematical physics literature, our inner products are conjugate-linear in the first argument and linear in the second. Also, for a Hermitian $X \in \mathcal{B}(\mathbb{C}^d)$, we write $X \geq 0$ if and only if X is positive semidefinite. We denote by $\mathbb{1}_{d_A} = \mathbb{1}_A \in \mathcal{B}(\mathbb{C}^{d_A})$, the identity operator, whereas we use the notation $\mathrm{id}_{d_A} = \mathrm{id}_A = \mathrm{id}_{\mathcal{B}(\mathbb{C}^{d_A})} \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}))$ for the identity map, that is, $\mathrm{id}_A(X) = X$ for any $X \in \mathcal{B}(\mathbb{C}^{d_A})$.

Moreover, as is standard in quantum theory, we use Dirac bra-ket notation. That is, a vector $\psi \in \mathbb{C}^d$ is denoted with a ket $|\psi\rangle$. The corresponding element of the dual space is denoted by a bra $\langle\psi|$, i.e., $\langle\psi| : \mathbb{C}^d \to \mathbb{C}$, $\mathbb{C}^d \ni |\varphi\rangle \mapsto \langle\psi|\varphi\rangle \coloneqq \langle\psi, \varphi\rangle$, where the inner product is the standard inner product on $\mathbb{C}^d$. Accordingly, if $|\psi\rangle \in \mathbb{C}^{d_B}$ and $|\varphi\rangle \in \mathbb{C}^{d_A}$, then we define $|\psi\rangle\langle\varphi| \in \mathcal{B}(\mathbb{C}^{d_A}; \mathbb{C}^{d_B})$ via the mapping $\mathbb{C}^{d_A} \ni |\phi\rangle \mapsto \langle\varphi|\phi\rangle\,|\psi\rangle \in \mathbb{C}^{d_B}$.

Our presentation in this chapter is structured as follows: Section 2.1 introduces the mathematical framework for describing quantum systems and measurements thereof. Section 2.2 concerns evolutions of quantum systems and their description in terms of completely positive maps. Finally,

in Section 2.3, we discuss so-called quantum superchannels and their connection to semicausal and semilocalizable quantum operations.

2.1 Quantum States and Measurements

First, we introduce quantum states and density matrices, which we use to describe quantum systems:

Definition 2.1.1 (Quantum States and Density Matrices)**.** *Let $d \in \mathbb{N}_{\geq 1}$. The set of d-dimensional* quantum states *or* density matrices *is*

$$\mathcal{S}(\mathbb{C}^d) := \left\{ \rho \in \mathcal{B}(\mathbb{C}^d) \mid \rho \geq 0 \ \wedge \ \mathrm{tr}[\rho] = 1 \right\} . \tag{2.1.1}$$

If $d = 2$, we often speak of *qubit* states, where qubit is an abbreviation for "quantum bit". Similarly, when $d \in \mathbb{N}_{\geq 1}$ is not specified, we will use the term *qudit*. Throughout this thesis, we will use the terms "quantum state" and "density matrix" effectively interchangeably.
At this point, we also make some additional remarks on $\mathcal{S}(\mathbb{C}^d)$ and the nomenclature surrounding it. First, the set $\mathcal{S}(\mathbb{C}^d)$ of d-dimensional density matrices is convex. Its extreme points are the rank-1 projections, which we call *pure* states. Any non-pure state is a *mixed* state. If $\rho \in \mathcal{S}(\mathbb{C}^d)$ is pure, then there is a unique $|\psi\rangle \in \mathbb{C}^d$ with $\langle\psi|\psi\rangle = 1$ such that $\rho = |\psi\rangle\langle\psi|$, and vice versa. Thus, we also often speak of normalized d-dimensional complex vectors as pure states, thereby implicitly identifying a rank-1 projection $|\psi\rangle\langle\psi|$ with the vector $|\psi\rangle$. Second, notice that $\mathcal{S}(\mathbb{C}^d)$ constitutes a generalization of the set of classical d-dimensional probability vectors. Namely, the diagonal $d \times d$-density matrices are in 1-to-1 correspondence with non-negative and normalized d-dimensional vectors, also known as probability vectors.
While we use quantum states to describe quantum systems, quantum theory tells us that these states are not directly accessible in experiments. Rather, we can only make observations about a quantum system by performing measurements on it. Therefore, we next describe our mathematical formalism for measurements:

Definition 2.1.2 (Measurements, Effect Operators, and Positive Operator-Valued Measures)**.** *Let $d \in \mathbb{N}_{\geq 1}$. The set of d-dimensional effect operators is*

$$\mathcal{E}(\mathbb{C}^d) := \left\{ E \in \mathcal{B}(\mathbb{C}^d) \mid 0 \leq E \leq \mathbb{1}_d \right\} . \tag{2.1.2}$$

An (n-outcome) positive operator-valued measure*, abbreviated as* POVM*, is a family $\{E_i\}_{i=1}^n$ of $n \in \mathbb{N}_{\geq 1}$ effect operators $E_i \in \mathcal{E}(\mathbb{C}^d)$ such that*

$$\sum_{i=1}^{n} E_i = \mathbb{1}_d . \tag{2.1.3}$$

We take a POVM $\{E_i\}_{i=1}^n$ as the mathematical description of a measurement with n possible outcomes. We will use the terms "POVM" and "measurement" synonymously. The so-called *Born rule* states that, when performing the measurement $\{E_i\}_{i=1}^n$ on a quantum system in the state $\rho \in \mathcal{S}(\mathbb{C}^d)$, the probability of observing the outcome $i \in \{1, \ldots, n\}$ is given by $p_i = \mathrm{tr}[\rho E_i]$.

Here, Definitions 2.1.1 and 2.1.2 ensure that the vector $(\mathrm{tr}[\rho E_i])_{i=1}^n$ indeed forms an n-dimensional probability vector. Moreover, quantum theory postulates that the state of the quantum system after we observe the measurement outcome i is the post-measurement state $\frac{1}{p_i}\sqrt{E_i}\rho\sqrt{E_i}$.

Composite Systems: Above, we have defined states and measurements for single quantum systems. Next, we consider composite systems. If system A is described by a Hilbert space $\mathbb{C}^{d_A}$, and system B has an associated Hilbert space $\mathbb{C}^{d_B}$, then the joint AB-system has the tensor product Hilbert space $\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B} \cong \mathbb{C}^{d_A \cdot d_B}$. Accordingly, states of the joint system are elements of $\mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$, effect operators are taken from $\mathcal{E}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$. Importantly, while $\mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ contains *product states* of the form $\rho_A \otimes \rho_B$, with $\rho_A \in \mathcal{S}(\mathbb{C}^{d_A})$ and $\rho_B \in \mathcal{S}(\mathbb{C}^{d_B})$, not all elements of $\mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ are of this form. Also *separable states*, given by convex combinations of product states, can be found in $\mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$. Crucially, however, $\mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ even contains non-separable states, which we call *entangled*. As an important example, we mention *maximally entangled states*: If $d_A = d_B = d$ and we pick an orthonormal basis $\{|a_i\rangle\}_{i=1}^d$ of $\mathbb{C}^{d_A} = \mathbb{C}^{d_B} = \mathbb{C}^d$, then we call the pure state $|\Omega\rangle := \frac{1}{\sqrt{d}} \sum_{i=1}^d |a_i\rangle \otimes |a_i\rangle$ the maximally entangled state with respect to the chosen orthonormal basis. The corresponding density matrix is $\Omega := |\Omega\rangle\langle\Omega| = \frac{1}{d} \sum_{i,j=1}^d |a_i\rangle\langle a_j| \otimes |a_i\rangle\langle a_j|$. Closely related to Ω is the so-called *SWAP operator*. For a general composite system, we define $\mathrm{SWAP} = \mathrm{SWAP}_{A,B} \in \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}; \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_A})$ by linearly extending $\mathrm{SWAP}\,|a_i\rangle \otimes |b_j\rangle = |b_j\rangle \otimes |a_i\rangle$, where $\{|a_i\rangle\}_{i=1}^{d_A}$ and $\{|b_i\rangle\}_{i=1}^{d_B}$ are (orthonormal) bases of $\mathbb{C}^{d_A}$ and $\mathbb{C}^{d_B}$, respectively. If $d_A = d_B = d$ and if we choose the same orthonormal basis $\{|a_i\rangle\}_{i=1}^d$ for both tensor factors and for the definition of the maximally entangled state, then SWAP and Ω are related through partial transposition. That is,

$$\mathrm{SWAP} = \sum_{i,j=1}^{d} |a_j\rangle\langle a_i| \otimes |a_i\rangle\langle a_j| = d\Omega^{T_A}\,, \tag{2.1.4}$$

where the *partial transpose* X^{T_A} of $X \in \mathcal{B}(\mathbb{C}^d \otimes \mathbb{C}^d)$ with respect to the chosen basis is defined via $\langle a_i| \otimes \langle a_j|\, X\, |a_k\rangle \otimes |a_\ell\rangle := \langle a_k| \otimes \langle a_j|\, X\, |a_i\rangle \otimes |a_\ell\rangle$. Unfortunately, an in-depth discussion of the relevance of quantum entanglement to the theory of quantum information is beyond the scope of this thesis. The interested reader is referred to [21, Chapter 4] for a more detailed introduction to this material.

An important tool in the study of composite quantum systems is the *partial trace*. We define *the partial trace over the B-subsystem* of a composite AB-system as the unique linear map $\mathrm{tr}_B : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A})$ that satisfies

$$\mathrm{tr}[\mathrm{tr}_B[X_{AB}]Y_A] = \mathrm{tr}[X_{AB}(Y_A \otimes \mathbb{1}_B)] \quad \text{for all } X_{AB} \in \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}),\ Y_A \in \mathcal{B}(\mathbb{C}^{d_A})\,. \tag{2.1.5}$$

The partial trace over the A-subsystem is defined analogously. For $\rho_{AB} \in \mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$, we call $\rho_A := \mathrm{tr}_B[\rho_{AB}]$ and $\rho_B := \mathrm{tr}_A[\rho_{AB}]$ the *reduced density matrices* or *reduced states* of ρ_{AB}. According to Eq. (2.1.5), performing a measurement $\{E_i\}_{i=1}^n$ on the reduced state ρ_A of ρ_{AB} leads to exactly the same outcome statistics as performing the measurement $\{E_i \otimes \mathbb{1}_B\}_{i=1}^n$ on the state ρ_{AB} of the composite system. In that sense, the partial trace allows us to focus on subsystems of a composite quantum system.

Distance Measures: We now shortly discuss some common distance measures in quantum information theory. Here, "distance measure" is used in an informal sense, only three of the quantities introduced below are actual "distances" in the sense of a metric.

First, and maybe most naturally, we can equip the set $\mathcal{S}(\mathbb{C}^d)$ with the (for convenience scaled) trace norm $\|\cdot\|_1/2$. This allows us to measure the difference between two quantum states $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$ via the *trace distance* $\|\rho-\sigma\|_1/2 = \mathrm{tr}[|\rho-\sigma|]/2$. In fact, this distance is not only intuitive from a mathematical perspective, it also has an operational interpretation: The maximal success probability in distinghuishing $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$, assuming that either of the two is prepared with probability $1/2$, by performing a 2-outcome measurement on a single copy of the unknown state is given by $\sup_{E \in \mathcal{E}(\mathbb{C}^d)} \mathrm{tr}[E(\rho-\sigma)] + \frac{1}{2} = \frac{\|\rho-\sigma\|_1+1}{2}$ (see, e.g., [17, Chapter 9] for a derivation).

A second important measure of similarity between quantum states is the *fidelity*. For states $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$, the fidelity is defined as $F(\rho,\sigma) := \mathrm{tr}[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}]$. Notice that the fidelity is symmetric and satisfies $0 \leq F(\rho,\sigma) \leq 1 = F(\rho,\rho)$ for all $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$. However, it does not satisfy a triangle inequality and is thus not a metric, in contrast to the trace distance. In the special case of pure states $|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi| \in \mathcal{S}(\mathbb{C}^d)$, the fidelity becomes the absolute value of the overlap of the corresponding vectors, $F(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = |\langle\psi|\phi\rangle|$. The trace distance and the quantum fidelity are closely related. Namely, the Fuchs–van de Graaf relations [23] state that, for any $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$, $1 - F(\rho,\sigma) \leq \|\cdot\|_1/2 \leq \sqrt{1-F(\rho,\sigma)^2}$.

Despite there being many other useful ways of determining distances between quantum states, we only discuss one more, namely the *quantum relative entropy*. The quantum relative entropy between two quantum states $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$ that satisfy $\mathrm{supp}(\rho) \cap \ker(\sigma) = \emptyset$ is defined to be $D(\rho\|\sigma) := \mathrm{tr}[\rho\log(\rho)] - \mathrm{tr}[\rho\log(\sigma)]$. Here, the logarithm is taken with base 2. We can rewrite the relative entropy using the *von Neumann entropy* $S(\rho)$, which is defined as $S(\rho) := -\mathrm{tr}[\rho\log(\rho)]$. With this, the relative entropy becomes $D(\rho\|\sigma) = -\mathrm{tr}[\rho\log(\sigma)] - S(\rho)$. If $\mathrm{supp}(\rho) \cap \ker(\sigma) \neq \emptyset$, then we define $D(\rho\|\sigma) := +\infty$. A useful result in quantum information is the non-negativity of the relative entropy, i.e., that $D(\rho\|\sigma) \geq 0$ holds for any two quantum states $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$. Moreover, for $\rho, \sigma \in \mathcal{S}(\mathbb{C}^d)$, $D(\rho\|\sigma) = 0$ implies $\rho = \sigma$ (see, e.g., [17, Theorem 11.7] for a proof). However, the quantum relative entropy is neither symmetric nor does it satisfy a triangle inequality, so it does not define a metric. Nevertheless, it is an important tool for comparing two quantum states. For example, when comparing a bipartite state $\rho_{AB} \in \mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ to $\rho_A \otimes \rho_B$, the tensor product of its reduced density matrices, we obtain a measure for the correlation between the A- and the B-system in the state ρ_{AB}. This defines the *quantum mutual information* $I(A:B)_\rho := D(\rho_{AB}\|\rho_A \otimes \rho_B)$.

2.2 Completely Positive Maps and Quantum Channels

Next, we consider valid transformations of quantum systems. To be consistent with the probabilistic interpretation of quantum theory given by the Born rule, we take such transformations to be linear. Moreover, we need to preserve the non-negativity of outcome probabilities of measurements. This is part of the motivation for the following:

Definition 2.2.1 (Completely Positive Maps)**.** *A linear map* $T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B})$ *is called* completely positive (CP) *if for any* $d_E \in \mathbb{N}_{\geq 1}$ *and for any* $\rho \in \mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_A})$ *with* $\rho \geq 0$ *also* $(\mathrm{id}_{\mathcal{B}(\mathbb{C}^{d_E})} \otimes T)(\rho) \geq 0$*. We denote the set of CP maps by*

$$\mathcal{CP}_{d_A,d_B} \coloneqq \left\{ T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B}) \mid T \text{ is linear and CP} \right\} . \tag{2.2.1}$$

If $d_A = d_B = d$*, we also write* $\mathcal{CP}_d \coloneqq \mathcal{CP}_{d,d}$*.*

Definition 2.2.1 captures the following intuition: Suppose A is our quantum system of interest and we consider an additional quantum system E of some arbitrary finite dimension. Now, suppose that only system A evolves non-trivially. Then, we have to preserve non-negativity of outcome probabilities for measurements performed on the A-system also when viewing it as a subsystem of the bipartite AE-system. In fact, it turns out (e.g., due to Theorem 2.2.3) that it suffices to consider auxiliary systems of dimension $d_E \leq d_A$ in the definition of complete positivity.
Now, adding a suitable normalization condition to the CP property, we arrive at the following notion of transformations of quantum systems:

Definition 2.2.2 (Quantum Channels and Operations – Schrödinger Picture)**.** *A* quantum channel in the Schrödinger picture *is a linear* completely positive and trace-preserving (CPTP) *map. That is, a linear map* $T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B})$ *is a quantum channel if* T *is CP and* $\mathrm{tr}[T(A)] = \mathrm{tr}[A]$ *holds for all* $A \in \mathcal{B}(\mathbb{C}^{d_A})$*. We denote the set of CPTP maps by*

$$\mathcal{CPTP}_{d_A,d_B} \coloneqq \left\{ T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B}) \mid T \text{ is linear and CPTP} \right\} . \tag{2.2.2}$$

If $d_A = d_B = d$*, we also write* $\mathcal{CPTP}_d \coloneqq \mathcal{CPTP}_{d,d}$*.*

Defintion 2.2.2 is formulated in the *Schrödinger picture*, where we think of quantum states as evolving and of measurements as remaining fixed. Namely, a quantum channel in the Schrödinger picture exactly maps quantum states to quantum states, even when embedding the evolving quantum system into a larger system.
We can, however, also take the view of the *Heisenberg picture*, i.e., of fixed states and evolving measurements. Mathematically, changing from the Schrödinger to the Heisenberg picture corresponds to taking the adjoint with respect to the Hilbert-Schmidt inner product. Namely, for a linear map $T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B})$ describing the evolution of states in the Schrödinger picture, we define the corresponding Heisenberg picture evolution $T^* : \mathcal{B}(\mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A})$ via the requirement

$$\mathrm{tr}[E^\dagger T(\rho)] = \mathrm{tr}[(T^*(E))^\dagger \rho] , \quad \text{for all } \rho \in \mathcal{B}(\mathbb{C}^{d_A}),\ E \in \mathcal{B}(\mathbb{C}^{d_B}) . \tag{2.2.3}$$

The Schrödinger picture map T is CP if and only if T^* is CP, and T is CPTP if and only if T^* is CP and unital (*CPU*). Here, T^* is called unital if $T^*(\mathbb{1}_B) = \mathbb{1}_A$. In analogy to the notation for CPTP maps introduced above, we will use $\mathcal{CPU}_{d_A,d_B}$ and $\mathcal{CPU}_d$ to denote sets of CPU maps.

Representations of Completely Positive Maps: We conclude our discussion of evolutions of quantum systems in terms of CPTP or CPU maps by presenting different useful character-

izations and representations of these maps. The first of these is a similarity transform that in particular facilitates checking complete positivity:

Theorem 2.2.3 (Choi-Jamiolkowski Isomorphism [24, 25]). *Fix an orthonormal basis $\{|a_i\rangle\}_i$ of $\mathbb{C}^{d_A}$. Consider the linear map*

$$\mathfrak{C}_{A;B} : \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B})) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}), \ \mathfrak{C}_{A;B}(T) = d(\mathrm{id}_A \otimes T)(|\Omega\rangle\langle\Omega|), \tag{2.2.4}$$

where $|\Omega\rangle := \frac{1}{\sqrt{d}} \sum_{i=1}^{d_A} |a_i\rangle \otimes |a_i\rangle$.

1. *$\mathfrak{C}_{A;B}$ is bijective, its inverse is given by $\mathfrak{C}_{A;B} : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$, $\mathfrak{C}_{A;B}^{-1}(\tau)(\rho) = \mathrm{tr}_A\left[(\rho^T \otimes \mathbb{1}_B)\tau\right]$. We call $\mathfrak{C}_{A;B}$ the* Choi-Jamiolkowski isomorphism.
2. *$T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$ is Hermiticity-preserving if and only if $\mathfrak{C}_{A;B}(T)$ is Hermitian.*
3. *$T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$ is CP if and only if $\mathfrak{C}_{A;B}(T) \geq 0$.*
4. *$T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$ is TP if and only if $\mathrm{tr}_B[\mathfrak{C}_{A;B}(T)] = \mathbb{1}_A$.*
5. *$T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$ is unital if and only if $\mathrm{tr}_A[\mathfrak{C}_{A;B}(T)] = \mathbb{1}_B$.*

Through the spectral decomposition of the (positive semidefinite) *Choi matrix* $\mathfrak{C}_{A;B}(T)$ of a CP map T, Theorem 2.2.3 gives rise to the following representation of CP maps:

Theorem 2.2.4 (Kraus Representation [26]). *A linear map $T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B})$ is CP if and only if there exist so-called* Kraus operators *$K_1, \ldots, K_r \in \mathcal{B}(\mathbb{C}^{d_A}; \mathbb{C}^{d_B})$, with $r \in \mathbb{N}_{\geq 1}$, such that*

$$T(A) = \sum_{i=1}^{r} K_i A K_i^\dagger . \tag{2.2.5}$$

In this case, T is CPTP if and only if $\sum_{i=1}^r K_i^\dagger K_i = \mathbb{1}_{d_A}$, and T is CPU if and only if $\sum_{i=1}^r K_i K_i^\dagger = \mathbb{1}_{d_B}$.

In particular, when deriving Theorem 2.2.4 from Theorem 2.2.3, we directly see that the number r of Kraus operators in Theorem 2.2.4 can be taken to be equal to the rank of the Choi matrix $\tau = \mathfrak{C}_{A;B}(T)$. Thus, we can in particular always find a Kraus representation with $r \leq d_A \cdot d_B$ Kraus operators. While the Kraus operators are not unique, any two sets of Kraus operators for the same CP map are unitarily related (see, e.g., [20, Theorem 2.1] for a more detailed statement and proof).
Moreover, when considering the linear map $V : \mathbb{C}^{d_A} \to \mathbb{C}^{d_B} \otimes \mathbb{C}^r$ defined by $V = \sum_{i=1}^r K_i \otimes |i\rangle$, where $\{|i\rangle\}_{i=1}^r$ is some orthonormal basis of $\mathbb{C}^r$, we obtain the following further useful representation for CP maps:

Theorem 2.2.5 (Stinespring Dilation – Heisenberg Picture [27]). *A linear map $T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B})$ is CP if and only if there exists a finite-dimensional auxiliary space $\mathbb{C}^{d_E}$ and a linear map $V : \mathbb{C}^{d_A} \to \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_E}$ such that*

$$T(X) = V^\dagger (X \otimes \mathbb{1}_E) V , \quad \text{for all } X \in \mathcal{B}(\mathbb{C}^{d_A}) . \tag{2.2.6}$$

Moreover, V is an isometry – i.e., V satisfies $V^\dagger V = \mathbb{1}_A$ – if and only if T is unital.

In the setting of Theorem 2.2.5, we call $\mathbb{C}^{d_E}$ a *dilation space* and V a *Stinespring isometry*. We will refer to any representation of a CP map in terms of a dilation space and a Stinespring isometry as *Stinespring dilation*.
Coming from the Kraus representation, it is easy to see that can find Stinespring dilations for any dilation space dimension $d_E \geq \operatorname{rank}(\mathfrak{C}_{A;B}(T))$. A Stinespring dilation with $d_E = \operatorname{rank}(\mathfrak{C}_{A;B}(T))$ is called *minimal*. Minimal Stinespring dilations are unique up to local unitaries acting on the dilation space and give rise to all possible dilations via local isometries on the dilation space (see, e.g., [20, Section 2.2] for a proof).
As Theorem 2.2.5 characterizes CP maps, we can apply it both in the Heisenberg and in the Schrödinger picture. It is convenient to think of the CP map T in Theorem 2.2.5 as describing the Heisenberg picture evolution, this then gives us the following representation for Schrödinger picture CPTP maps:

Theorem 2.2.6 (Stinespring Dilation– Schrödinger Picture)**.** *A linear map $T : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_B})$ is CPTP if and only if there exist a unitary $U \in \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_B})$ and a pure state $|\varphi_0\rangle \in \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_B}$, such that*

$$T(\rho) = \operatorname{tr}_{AB}[U(\rho \otimes |\varphi_0\rangle\langle\varphi_0|)U^\dagger], \quad \text{for all } \rho \in \mathcal{B}(\mathbb{C}^{d_A}), \tag{2.2.7}$$

where the partial trace is over the first two tensor factors of $\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_B}$.

Theorem 2.2.6 is also often referred to as giving an *open system representation* for CPTP maps. Namely, if T is a linear CPTP map, we can understand T as the reduced map of a unitary acting on an intially uncorrelated composite system.
With these different characterizations of CP, CPTP, and CPU maps at hand, we can now easily construct concrete examples. We want to highlight just three such examples.

Example 2.2.7 (Unitary Channels)**.** In standard quantum mechanics, an evolution of a quantum system is described by a unitary. While this is not the most general kind of evolution allowed in our framework, we do recover it as a special case. Namely, if $U \in \mathcal{B}(\mathbb{C}^d)$ is a unitary, then the linear map $\mathcal{B}(\mathbb{C}^d) \ni \rho \mapsto U\rho U^\dagger$ is CPTP by Theorem 2.2.4 and thus describes a valid transformation of a quantum system. We call such a CPTP map a *unitary quantum channel*. In addition to being CPTP, a unitary quantum channel maps pure states to pure states and is invertible, with its inverse given by the unitary quantum channel with unitary $U^\dagger$.

Example 2.2.8 (Partial Trace)**.** In Section 2.1, we have encountered the partial trace. In fact, the partial trace $\operatorname{tr}_B : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A})$ as defined in Eq. (2.1.5) is a CPTP map. We can, e.g., see that as a consequence of Theorem 2.2.6, when taking a pure product state $|\varphi_0\rangle = |0\rangle \otimes |0\rangle \in \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_B}$ and as unitary $U \in \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_B} \otimes \mathbb{C}^{d_B})$ the swap operation between the second and the fourth tensor factor.

Example 2.2.9 (Measurement Channels)**.** Also measurements in quantum theory can be understood in the framework of CPTP maps. Namely, if we measure a POVM $\{E_i\}_{i=1}^n$, without recording the measurement outcome, the evolution of the quantum system through the measurement process is described by the CPTP map with Kraus representation $\sum_{i=1}^n \sqrt{E_i}(\cdot)\sqrt{E_i}$.

2.3 Superchannels, Semicausality, and Semilocalizability

After introducing quantum states as describing quantum systems in Section 2.1, we have considered quantum evolutions as CP maps on states in Section 2.2. Now, we take a step further by looking at a higher-order level of evolutions, namely evolutions of CP maps:

Definition 2.3.1 (Quantum Superchannels [28]). *A* quantum superchannel *is a linear map* $\hat{S} : \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B})) \to \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$ *such that, for any* $d_E \in \mathbb{N}_{\geq 1}$*, the map* $\hat{S}_{d_E}$ *defined as* $\hat{S}_{d_E} = \mathrm{id}_{\mathcal{B}(\mathcal{B}(\mathbb{C}^{d_E}))} \otimes \hat{S}$ *satisfies that*

(i) $\hat{S}_{d_E}(T)$ *is CP whenever* $T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_B}))$ *is CP, and*

(ii) $\hat{S}_{d_E}(T)$ *is a quantum channel whenever* $T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_B}))$ *is a quantum channel.*

Physically, the motivation behind Definition 2.3.1 is clear: Quantum superchannels take a valid quantum evolution as input and output another valid quantum evolution. From a computing perspective, quantum superchannels are a mathematical framework for describing quantum circuit boards with a free slot, into which we can then plug a gate to change the overall functionality. Studying quantum superchannels becomes feasible due to their close connection to quantum operations with a certain causal structure. We now introduce this causality assumption and then, in Theorem 2.3.5, discuss their connection to quantum superchannels:

Definition 2.3.2 (Semicausal CP Maps – Heisenberg picture [29, 30]). *A linear CP map* $T : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ *in the Heisenberg picture is* Heisenberg $B \not\to A$-semicausal *if there exists a linear CP map* $T^A : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_A})$ *such that*

$$T(X \otimes \mathbb{1}_B) = T^A(X) \otimes \mathbb{1}_B\,, \qquad \text{for all } X \in \mathcal{B}(\mathbb{C}^{d_A})\,. \tag{2.3.1}$$

When changing from the Heisenberg to the Schrödinger picture, this leads us to define: A CP map $T_* : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ is called *Schrödinger* $B \not\to A$ *semicausal* if there exists a CP map $T_*^A : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_A})$ such that, for every $\rho_{AB} \in \mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$, $\mathrm{tr}_B[T(\rho_{AB})] = T_*^A(\mathrm{tr}_B[\rho_{AB}])$.
In the Schrödinger picture, the physical motivation for Definition 2.3.2 becomes clear: For a semicausal evolution of a bipartite system, the evolution of the A-subsystem must be independent of the input on the B-subsystem. Informally, this means that no information flow from the B- to the A-subsystem is allowed. The following definition introduces a natural, operationally motivated class of maps that satisfy this requirement:

Definition 2.3.3 (Semilocalizable Quantum Channels [29]). *A linear CP map* T *on a bipartite space,* $T : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$*, is called* Heisenberg $B \not\to A$ semilocalizable *if there exist a finite-dimensional auxiliary space* $\mathbb{C}^{d_E}$*, a linear CPU map* $F : \mathcal{B}(\mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_B})$*, and a linear CP map* $G : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_E}) \to \mathcal{B}(\mathbb{C}^{d_A})$ *such that*

$$T = (G \otimes \mathrm{id}_B)(\mathrm{id}_A \otimes F)\,. \tag{2.3.2}$$

Again, the corresponding notion in the Schrödinger picture is easily obtained by considering the adjoint with respect to the Hilbert-Schmidt inner product. Namely, a linear CP map $T_* : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ is Schrödinger $B \not\to A$ semilocalizable if and only if we can write $T_* = (\mathrm{id}_A \otimes F_*)(G_* \otimes \mathrm{id}_B)$ for some linear CPTP map $F_* : \mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_B})$ and some linear CP map $G_* : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_E})$.
Again, the Schrödinger picture perspective allows us to easily interpret Definition 2.3.3. Namely, a semilocalizable evolution is implemented as follows: First, act on the A-system to produce a bipartite output on the AE-system. Then, transmit the E-subsystem of the output to B. Finally, act on the joint EB-system to produce the B-system output. Intuitively, such a $B \not\to A$ semilocalizable procedure in particular does not allow for information flow from the B- to the A-system and is thus $B \not\to A$ semicausal. This can be easily checked by verifying that a CP map satisfying Eq. (2.3.2) also satisfies Eq. (2.3.1). An important insight into semicausality and semilocalizability is that the converse also holds for CP maps:

Theorem 2.3.4 (Semicausality versus Semilocalizability [30]). *A CP map $T : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ is Heisenberg $B \not\to A$ semilocalizable if and only if it is Heisenberg $B \not\to A$ semicausal.*

This result was first proved by [30], later reproved with a different reasoning by [31], and extended to infinite dimensions in [8]. Theorem 2.3.4 tells us that any semicausal CP map can be realized in a semilocalizable manner. In particular, the operational interpretation of semilocalizability carries over to semicausality. In fact, semicausal (and thus semilocalizable) CP maps are closely connected to quantum superchannels:

Theorem 2.3.5 (Quantum Superchannels versus Semicausal CP Maps [28]). *Consider a linear map $\hat{S} : \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B})) \to \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$, write $S : \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}) \to \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$, $S \coloneqq \mathfrak{C}_{A;B} \circ \hat{S} \circ \mathfrak{C}_{A;B}^{-1}$. $\hat{S}$ is a quantum superchannel if and only if S is a Schrödinger $B \not\to A$ semicausal CP map whose reduced map $S^A : \mathcal{B}(\mathbb{C}^{d_A}) \to \mathcal{B}(\mathbb{C}^{d_A})$ satisfies $S^A(\mathbb{1}_A) = \mathbb{1}_A$.*

Remark 2.3.6. In this chapter, we have restricted our presentation to a mathematical formulation of quantum mechanics in finite dimensions. As already finite-dimensional quantum system offer a plethora of fascinating phenomena central to quantum theory, this restriction is widespread in quantum information theory. In particular, the Core Articles I-IV [1–4] as well as the Articles V [5] and VII [6] fall into this finite-dimensional framework. Among the contributed articles that investigate questions in quantum computing and information, only Article VIII [8] contains results for the infinite-dimensional case. Therefore, to enhance the clarity of the presentation, this chapter introduces only the tools of finite-dimensional theory. In Section III of Article VIII [8], we discuss the infinite-dimensional case in more detail.

While there are many further interesting features of quantum information theory, the short introduction given in this chapter already covers the aspects most relevant to the remainder of this thesis. Thus, this concludes our presentation of the mathematics behind quantum information theory.

Chapter 3

Mathematical Ingredients of Statistical Learning Theory

Having introduced our mathematical framework for quantum theory in Chapter 2, we devote this chapter to the second theory underlying this thesis, namely to *statistical learning theory*. Here, we focus on the specific aspects relevant to the remainder of this thesis and recommend textbooks such as [32–35] or lectures notes such as [36] as references for this and further material. Our presentation is structured as follows: In Section 3.1, we introduce the model of probably approximately correct learning, emphasizing the importance of generalization bounds. Section 3.2 reviews different complexity measures for function classes and the corresponding generalization guarantees. This is followed by an overview over relations between different complexity measures in Section 3.3. Finally, Section 3.4 discusses some aspects of undecidable problems in learning theory.

3.1 Probably Approximately Correct Learning

Our focus is on the standard statistical framework for learning problems with labelled data, so-called *supervised learning problems*. In this formalism, we consider a data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where we think of $\mathcal{X}$ as an input/instance space and of $\mathcal{Y}$ as an output/label space. Usually, if $\mathcal{Y}$ is discrete, we speak of a *classification* task, and if $\mathcal{Y}$ is continuous, we speak of a *regression* task. A *training data set* S of size m is a multiset $S = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\} = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\}$ consisting of m training examples $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. A *learning algorithm* $\mathcal{A}$ takes such a training data set as input and outputs a hypothesis in $\mathcal{Y}^{\mathcal{X}}$. That is, a learner $\mathcal{A}$ is a map

$$\mathcal{A} : \bigcup_{m \in \mathbb{N}_{\geq 1}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{Y}^{\mathcal{X}}, \; S \mapsto \mathcal{A}(S) = h_S \,. \tag{3.1.1}$$

Note that at this point, the "algorithm" in "learning algorithm" is not to be taken literally, here we do not assume $\mathcal{A}$ to be a (Turing) computable function. Informally, this means that we do

not demand there to be a computer program for evaluating $\mathcal{A}$. We refer to Section 3.4 for a discussion of select aspects computability in learning theory. We will call the class

$$\mathcal{F} \coloneqq \mathcal{A}\left(\bigcup_{m\in\mathbb{N}_{\geq 1}} (\mathcal{X}\times\mathcal{Y})^m\right) \subseteq \mathcal{Y}^{\mathcal{X}} \tag{3.1.2}$$

of all functions that the learning algorithm can produce as output hypothesis the *concept/hypothesis class* associated with $\mathcal{A}$.

Next, we introduce the underlying statistical assumption for our framework. Namely, we suppose that there is some probability measure P on $\mathcal{X}\times\mathcal{Y}$ such that the training examples are drawn independently and identically distributed (i.i.d.) according to P. Thus, we also refer to P as the *data-generating distribution/measure*. Crucially, we usually think of P as being unknown to the learner. Here and throughout, we tacitly assume that the corresponding σ-algebra is chosen as a product of Borel σ-algebras, and that all involved functions are suitably measurable. Also, we denote by $\mathrm{Prob}(\mathcal{X}\times\mathcal{Y})$ the set of all probability distributions on $\mathcal{X}\times\mathcal{Y}$, $\mathrm{Prob}(\mathcal{X})$ is used analogously.

To evaluate a learner's performance, we take a loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$, the choice of which should be adapted to the problem at hand. Intuitively, a large (small) value $\ell(y_1, y_2)$ indicates that $y_1 \in \mathcal{Y}$ and $y_2 \in \mathcal{Y}$ are far apart (close). For fairness, we assume that the learner knows the loss function ℓ according to which the performance is judged. Now, we can formulate the goal of the learner: Given access to training data, a learner should output a hypothesis $h \in \mathcal{F}$ that achieves a small *expected/true risk*

$$R(h) \coloneqq R_P(h) \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(\boldsymbol{x}))\, \mathrm{d}P(\boldsymbol{x}, y)\,. \tag{3.1.3}$$

Note that the true risk $R(h)$ depends on both the loss function ℓ and the probability measure P. To illustrate this definition, we consider two commonly used loss functions. If $\mathcal{Y} = \{1, \ldots, k\}$ is a discrete space of $k \in \mathbb{N}_{>1}$ labels, one often uses the 0-1-loss defined as $\ell(y_1, y_2) \coloneqq 1 - \delta_{y_1,y_2}$, with δ_{y_1,y_2} denoting the Kronecker delta. This leads to the probability of misclassification as expected risk $R(h) = \mathbb{P}_{(\boldsymbol{x},y)\sim P}[h(\boldsymbol{x}) \neq y]$. For a continuous target space $\mathcal{Y} = \mathbb{R}$, if we take the square loss $\ell(y_1, y_2) \coloneqq (y_1 - y_2)^2$, the expected risk becomes the mean squared error $R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}[(y - h(\boldsymbol{x}))^2]$.

The notion of true risk now allows us to formulate what we mean by learning:

Definition 3.1.1 (Agnostic PAC Learning [37, 38]). *A learner $\mathcal{A} : \bigcup_{m\in\mathbb{N}_{\geq 1}} (\mathcal{X}\times\mathcal{Y})^m \to \mathcal{Y}^{\mathcal{X}}$, $S \mapsto \mathcal{A}(S) = h_S$ with hypothesis class $\mathcal{F} \coloneqq \mathcal{A}\left(\bigcup_{m\in\mathbb{N}_{\geq 1}} (\mathcal{X}\times\mathcal{Y})^m\right) \subseteq \mathcal{Y}^{\mathcal{X}}$ is an* agnostic (ε, δ)-probably approximately correct ((ε, δ)-PAC) *learner for $\mathcal{F}$, where $\varepsilon, \delta \in (0, 1)$, from training data of size $m \geq m_0 = m_0(\varepsilon, \delta) \in \mathbb{N}_{\geq 1}$ if the following holds: For every $P \in \mathrm{Prob}(\mathcal{X}\times\mathcal{Y})$, with probability $\geq 1 - \delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,*

$$R(h_S) \leq \inf_{h\in\mathcal{H}} R(h) + \varepsilon\,. \tag{3.1.4}$$

Moreover, we say that $\mathcal{A}$ *is an* agnostic PAC learner *for* $\mathcal{F}$ *if there exists a map* $m_{\mathcal{F}} : (0,1) \times (0,1) \to \mathbb{N}_{\geq 1}$ *such that* $\mathcal{A}$ *is an agnostic* (ε, δ)*-PAC learner for* $\mathcal{F}$ *from training data of size* $m \geq m_{\mathcal{F}}(\varepsilon, \delta)$ *for all* $\varepsilon, \delta \in (0,1)$*. Finally, we say that* $\mathcal{F}$ *is* agnostically PAC learnable *if there exists an agnostic PAC learner* $\mathcal{A}$ *for* $\mathcal{F}$*.*

While our focus mostly will be on the setting of Definition 3.1.1, there is a related model that is also often of interest:

Definition 3.1.2 (Realizable PAC Learning [39, 40])**.** *A learner* $\mathcal{A} : \bigcup_{m \in \mathbb{N}_{\geq 1}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{Y}^{\mathcal{X}}$, $S \mapsto \mathcal{A}(S) = h_S$ *is a* realizable (ε, δ)-PAC *learner for a concept class* $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$*, where* $\varepsilon, \delta \in (0,1)$*, from training data of size* $m \geq m_0 = m_0(\varepsilon, \delta) \in \mathbb{N}_{\geq 1}$ *if the following holds: For every* $P \in \mathrm{Prob}(\mathcal{X})$ *and for every* $f_* \in \mathcal{F}$*, with probability* $\geq 1 - \delta$ *over the choice of a training data set* $S = \{(\boldsymbol{x}_i, f_*(\boldsymbol{x}_i))\}_{i=1}^m$ *consisting of* m *examples, with the* $\boldsymbol{x}_i$ *drawn i.i.d. according to* P*,*

$$R(h_S) \leq \varepsilon\,. \tag{3.1.5}$$

Moreover, we say that $\mathcal{A}$ *is a* realizable PAC learner *for* $\mathcal{F}$ *if there exists a map* $m_{\mathcal{F}} : (0,1) \times (0,1) \to \mathbb{N}_{\geq 1}$ *such that* $\mathcal{A}$ *is a realizable* (ε, δ)*-PAC learner for* $\mathcal{F}$ *from training data of size* $m \geq m_{\mathcal{F}}(\varepsilon, \delta)$ *for all* $\varepsilon, \delta \in (0,1)$*. Finally, we say that* $\mathcal{F}$ *is* realizably PAC learnable *if there exists a realizable PAC learner* $\mathcal{A}$ *for* $\mathcal{F}$*.*

It is immediate from these two definitions that agnostic PAC learning describes a more general scenario than realizable PAC learning. In fact, the realizable case goes back to [39, 40], the agnostic case was then introduced later by [37, 38]. We will phrase most of the results in this chapter in the agnostic framework and only occasionally comment on possible strengthenings under the realizability assumption, i.e., under the assumption that there is some hypothesis in $\mathcal{F}$ that achieves zero error.

We also note that Definition 3.1.2 describes what is often referred to as *improper* (or *representation-independent*) realizable PAC learning. Here, we speak of improper learning because we do not require the learner $\mathcal{A}$ to output a hypothesis from the class $\mathcal{F}$ with respect to which the data is assumed to be realizable. If we change the definition to also restrict ourselves to learners whose range is contained in $\mathcal{F}$, we obtain the notion of *proper* realizable PAC learning.

With Definitions 3.1.1 and 3.1.2, we have formalized what we mean by successful learning. Next, we investigate how this can be achieved. The following observation is crucial for our discussion: A learner that has access to a training data set S does not have sufficient information to evaluate the true risk since she does not know the data-generating distribution P. A natural approach towards overcoming this challenge is to employ the training data to build a proxy for the true risk. That is, given a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ of size m, we define the *empirical/training risk* of a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ to be

$$\hat{R}_S(h) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i))\,. \tag{3.1.6}$$

In contrast to the true risk of Eq. (3.1.3), a learner that knows the training data set S, the loss function ℓ, and the hypothesis h can indeed evaluate the empirical risk of Eq. (3.1.6). Thus,

empirical risk minimization (ERM) provides a natural way of approaching the learning problem: Given training data, we attempt to find a hypothesis that achieves small (or even minimal) empirical risk.

To prepare the statistical analysis of approaches based on the empirical risk, we now consider the so-called *estimation error* of ERM. Namely, given training data S, we will denote by $\hat{h} = \hat{h}(S) \in \mathcal{F}$ a hypothesis that minimizes the empirical risk among hypotheses in $\mathcal{F}$. We define the estimation error of $\hat{h}$ to be

$$R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h)\,, \tag{3.1.7}$$

the difference between the true risk achieved by $\hat{h}$ and the optimal true risk achievable by a function in $\mathcal{F}$. By Definition 3.1.1, for ERM to be a valid agnostic PAC learner, we want exactly this expression to be small, with high probability. We now insert a zero into Eq. (3.1.7) and use the defining property of $\hat{h}$ to observe that we can bound the estimation error as

$$R(\hat{h}) - \inf_{h \in \mathcal{F}} R(h) = \left(R(\hat{h}) - \hat{R}_S(\hat{h})\right) + \sup_{h \in \mathcal{F}} \left(\hat{R}_S(\hat{h}) - R(h)\right) \tag{3.1.8}$$

$$\leq 2 \sup_{h \in \mathcal{F}} |R(h) - \hat{R}(h)|\,. \tag{3.1.9}$$

For a hypothesis h, the difference $R(h) - \hat{R}(h)$ is the *generalization error* of h. Thus, the above inequality tells us: We can control the estimation error of ERM if we have uniform (over the hypothesis class $\mathcal{F}$) bounds on the absolute generalization error.

Motivated by this discussion, we extract the aspect of generalization in PAC learning:

Definition 3.1.3 (PAC Generalization Bounds)**.** *Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. A* PAC generalization bound *for $\mathcal{F}$ is a guarantee of the following form: For every probability distribution P over $\mathcal{X} \times \mathcal{Y}$ and for every $\delta \in (0,1)$, with probability $\geq 1 - \delta$ over the choice of training data $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,*

$$\forall h \in \mathcal{F}:\ |R(h) - \hat{R}(h)| \leq g_{\mathcal{F}}(m, \delta, h, P)\,, \tag{3.1.10}$$

where $g_{\mathcal{F}} : \mathbb{N}_{\geq 1} \times (0,1) \times \mathcal{F} \times \mathrm{Prob}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_{\geq 0}$, with $\mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ denoting the set of all probability distributions over $\mathcal{X} \times \mathcal{Y}$.

Naturally, for Eq. (3.1.10) to be useful, we aim for an upper bound that in particular satisfies $\lim_{m \to \infty} g_{\mathcal{F}}(m, \delta, h, P) = 0$, and we are interested in the speed of convergence. In the setting of Definition 3.1.3, if we have Eq. (3.1.10) with a P-independent upper-bound $g_{\mathcal{F}}(m, \delta, h)$, where $g_{\mathcal{F}} : \mathbb{N}_{\geq 1} \times (0,1) \times \mathcal{F} \to \mathbb{R}_{\geq 0}$, we speak of a *distribution-independent* generalization bound. Also, if Eq. (3.1.10) holds with an h-independent bound $g(m, \delta, P)$, with $g_{\mathcal{F}} : \mathbb{N}_{\geq 1} \times (0,1) \times \mathrm{Prob}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_{\geq 0}$, we call the generalization bound *uniform (over $\mathcal{F}$)*. By the above discussion, distribution-independent uniform generalization bounds for $\mathcal{F}$ can be used to give PAC learning guarantees for ERM in the sense of Definition 3.1.1.

The remainder of this chapter, as well as much of this thesis, focuses on deriving generalization bounds. Notice, however, that generalization is not the only important aspect in machine learning

problems. For instance, there is also an often significant optimization challenge in finding a hypothesis that (approximately) minimizes the empirical risk, given a data set. Moreover, also the choice of a machine learning model, which then dictates the hypothesis class $\mathcal{F}$ and thus the optimal achievable risk $\inf_{h\in\mathcal{F}} R(h)$, is crucial. While these aspects are certainly important, discussing them in detail exceeds the scope of this thesis. Rather, we emphasize generalization as the statistical aspect at the heart of PAC learning.

Remark 3.1.4. In this section, we have considered deterministic learning procedures. The scenario can be extended to allow for stochasticity in the learning algorithm. Namely, if we consider a learner $\mathcal{A}$ that, upon input of a training data set S, outputs a probability measure μ_S over output hypotheses, we can study the expected true risk $\mathbb{E}_{h\sim\mu_S}[R(h)]$ and the expected empirical risk $\mathbb{E}_{h\sim\mu_S}[\hat{R}_S(h)]$, where the expectations are with respect to hypotheses drawn from μ_S. This is the perspective usually taken in the PAC-Bayesian framework [41]. Alternatively, for randomized learning algorithms, we can also consider variants of Definitions 3.1.1,3.1.2, and 3.1.3 in which we require the respective statements to hold with high probability over the randomness both in the choice of the data and in the learner itself.

Remark 3.1.5. There are other mathematical frameworks for formalizing tasks of learning from data. Beyond the PAC framework, some influential models include learning from membership queries or equivalence queries [42], learning from statistical queries [43], the mistake bound model of online learning [44], regret minimization in online learning [45], and different scenarios for teacher-learner interactions [46–48].

3.2 Complexity Measures and Generalization Bounds

A well established path towards PAC generalization bounds in classical learning theory leads through complexity measures for concept classes. In this section, we review some of those complexity measures and the resulting generalization bounds. We begin with a combinatorial dimension for binary-valued function classes:

Definition 3.2.1 (VC Dimension [39])**.** *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$. A set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\} \subseteq \mathcal{X}$ is* shattered *by $\mathcal{F}$ if for any $C \subseteq \{1, ..., k\}$ there exists a function $f_C \in \mathcal{F}$ such that for all $1 \leq i \leq k$, $i \in C$ if and only if $f_C(\boldsymbol{x}_i) = 1$. We define the* Vapnik-Chervonenkis (VC) dimension *of $\mathcal{F}$ as the largest size of a set shattered by $\mathcal{F}$:*

$$\mathrm{VCdim}(\mathcal{F}) := \sup\{k \in \mathbb{N}_0 \mid \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathcal{X} \text{ s.t. } \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \text{ is shattered by } \mathcal{F}\}\,. \tag{3.2.1}$$

Already from Definition 3.2.1, it makes intuitive sense to consider the VC dimension as measuring the complexity of a $\{0,1\}$-valued concept class. This intuition is strengthened when evaluating the VC dimension of simple geometric hypotheses classes, for example arising from axis-aligned rectangles [34, Example 3.14] or from more general convex polygons [34, Example 3.15], and of hypothesis classes obtained by post-processing elements of a real function vector space by the sign function [36, Theorem 1.9]. The following theorem demonstrates that the VC dimension indeed is a complexity measure useful for PAC generalization bounds:

Theorem 3.2.2 (Generalization Bound via VC Dimension [49, 50]). *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ and let $\ell : \{0,1\} \times \{0,1\} \to \{0,1\}$ be the 0-1-loss defined as $\ell(y_1, y_2) \coloneqq 1 - \delta_{y_1,y_2}$. Then, for every $P \in \mathrm{Prob}(\mathcal{X} \times \{0,1\})$ and for every $\delta \in (0,1)$, with probability $\geq 1 - \delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,*

$$\sup_{h \in \mathcal{F}} |R(h) - \hat{R}_S(h)| \leq C \cdot \sqrt{\frac{\mathrm{VCdim}(\mathcal{F}) + \ln(1/\delta)}{m}}\,, \tag{3.2.2}$$

where $C > 0$ is some universal constant.

Theorem 3.2.2 is paradigmatic for the – in this case, distribution-independent and uniform – PAC generalization bounds that we can derive from complexity measures. Any such generalization bound also comes with a corresponding sample complexity bound. In this case, it takes the following form:

Corollary 3.2.3 (Sample Complexity Bound via VC Dimension). *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, assume that $\mathrm{VCdim}(\mathcal{F}) < \infty$. Let $\ell : \{0,1\} \times \{0,1\} \to \{0,1\}$ be the 0-1-loss defined as $\ell(y_1, y_2) \coloneqq 1 - \delta_{y_1,y_2}$. Then, for every $P \in \mathrm{Prob}(\mathcal{X} \times \{0,1\})$ and for every $\varepsilon, \delta \in (0,1)$, a sample size*

$$m = m(\varepsilon, \delta) = C \cdot \frac{\mathrm{VCdim}(\mathcal{F}) + \ln(1/\delta)}{\varepsilon^2}\,, \tag{3.2.3}$$

where $C > 0$ is a universal constant, suffices to guarantee: With probability $\geq 1 - \delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,

$$\sup_{h \in \mathcal{F}} |R(h) - \hat{R}_S(h)| \leq \varepsilon\,. \tag{3.2.4}$$

As a consequence of the discussion in Section 3.1, Corollary 3.2.3 in particular implies that the sample size in Eq. (3.2.3) is sufficient for ERM to be an agnostic (ε, δ)-PAC learner for $\mathcal{F}$ in the sense of Definition 3.1.1. In the realizable case, this sample complexity guarantee can be improved with respect to the dependence on the accuracy ε. Namely, a sample size of

$$m = m(\varepsilon, \delta) = C \cdot \frac{\mathrm{VCdim}(\mathcal{F}) \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon} \tag{3.2.5}$$

suffices for ERM to be a (proper) realizable PAC learner for $\mathcal{F}$. In fact, if we allow for improper learning, the $\ln(1/\varepsilon)$ can be removed [51].
Interestingly, also sample complexity lower bounds that match these upper bounds up to constant factors are known [52–54], both for the agnostic and the realizable case. Thus, the VC dimension is central in understanding and quantifying the information-theoretic complexity of PAC binary classification. Naturally, this led to attempts at generalizing the VC dimension to other learning scenarios. Among the results of these efforts is the following combinatorial complexity measure for $\mathbb{R}$-valued concept classes:

Definition 3.2.4 (Pseudo-Dimension [55]). *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. A set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\} \subseteq \mathcal{X}$ is* pseudo-shattered *by $\mathcal{F}$ if there exist $y_1, ..., y_k \in \mathbb{R}$ such that for any $C \subseteq \{1, ..., k\}$ there exists a function*

$f_C \in \mathcal{F}$ such that for all $1 \leq i \leq k$, $i \in C$ if and only if $f_C(\boldsymbol{x}_i) \geq y_i$. We define the pseudo-dimension *of $\mathcal{F}$ as the largest size of a set pseudo-shattered by $\mathcal{F}$:*

$$\mathrm{Pdim}(\mathcal{F}) \coloneqq \sup\{k \in \mathbb{N}_0 \mid \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathcal{X} \text{ s.t. } \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \text{ is pseudo-shattered by } \mathcal{F}\}. \tag{3.2.6}$$

For a $\{0,1\}$-valued function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, we have $\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\mathcal{F})$, pseudo- and VC dimension coincide. Thus, the pseudo-dimension indeed is a generalization of the VC dimension. A sometimes useful refinement of the pseudo-dimension is provided in the following:

Definition 3.2.5 (Fat-Shattering Dimension [56])**.** *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and let $\alpha > 0$. A set $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\} \subseteq \mathcal{X}$ is* α-fat-shattered *by $\mathcal{F}$ if there exist $y_1, ..., y_k \in \mathbb{R}$ such that for any $C \subseteq \{1, ..., k\}$ there exists a function $f_C \in \mathcal{F}$ such that for all $1 \leq i \leq k$:*

1. *$i \notin C \Rightarrow f_C(\boldsymbol{x}_i) \leq y_i - \alpha$ and*
2. *$i \in C \Rightarrow f_C(\boldsymbol{x}_i) \geq y_i + \alpha$.*

The α-fat-shattering dimension *of $\mathcal{F}$ is defined to be*

$$\mathrm{fat}(\mathcal{F}, \alpha) \coloneqq \sup\{k \in \mathbb{N}_0 \mid \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathcal{X} \text{ s.t. } \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \text{ is } \alpha\text{-fat-shattered by } \mathcal{F}\}. \tag{3.2.7}$$

Trivially, $\mathrm{fat}(\mathcal{F}, \alpha) \leq \mathrm{fat}(\mathcal{F}, \beta) \leq \mathrm{Pdim}(\mathcal{F})$ holds for any $0 < \beta < \alpha$ and for any function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. In that sense, the fat-shattering dimension constitutes a refinement of the pseudo-dimension that includes a margin parameter.
Similarly to the VC dimension, also the fat-shattering and thus the pseudo-dimension can be used to obtain distribution-independent uniform generalization error bounds. We do not state these generalization bounds at this point because they can be obtained as corollaries of the remaining results in this section when combined with the insights of Section 3.3. For fat-shattering dimension-based generalization guarantees in the realizable case, see, e.g., [57, Corollary 3.3].
The combinatorial dimensions introduced above serve to quantify the complexity of a function class. In a learning problem, however, the functions and concept classes do not appear in isolation, but together with training data and a data-generating distribution. This observation motivates the study of distribution-dependent generalization guarantees, arising from data- or distribution-dependent complexity measures. Among those, a particularly prominent one is the Rademacher complexity:

Definition 3.2.6 ((Empirical) Rademacher Complexities [58])**.** *Let $\mathcal{Z}$ be a data space, let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ be an $\mathbb{R}$-valued function class, and let $S = \{\boldsymbol{z}_i\}_{i=1}^m \in \mathcal{Z}^m$ be a data set consisting of $m \in \mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i \in \mathcal{Z}$. The* empirical Rademacher complexity *of $\mathcal{H}$ with respect to S is defined as*

$$\hat{\mathcal{R}}_S(\mathcal{H}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim U(\{-1,1\}^m)} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\boldsymbol{z}_i) \right], \tag{3.2.8}$$

where $U(\{-1,1\}^m)$ denotes the uniform distribution on the Boolean hypercube $\{-1,1\}^m$. The i.i.d. random variables $\sigma_1, \ldots, \sigma_m$ are often called Rademacher random variables.

If $P \in \mathrm{Prob}(\mathcal{Z})$*, then the* Rademacher complexities *of* $\mathcal{H}$ *with respect to* P *are defined as*

$$\mathcal{R}_m(\mathcal{H}) \coloneqq \mathbb{E}_{S\sim P^m}\left[\hat{\mathcal{R}}_S(\mathcal{H})\right] \qquad \text{for } m \in \mathbb{N}_{\geq 1}\,, \tag{3.2.9}$$

where $S \sim P^m$ *means that* $S = \{\boldsymbol{z}_i\}_{i=1}^m$ *with the* $\boldsymbol{z}_i$ *drawn i.i.d. according to* P*.*

As a consequence of McDiarmid's bounded differences inequality [59], the empirical Rademacher complexity concentrates strongly around the Rademacher complexity with respect to the measure that the data is generated from. Thus, while we phrase the results in this section in terms of empirical Rademacher complexities, we can usually replace them by Rademacher complexities without changing the essential features of the results.
From the perspective of high-dimensional probability and random process theory, it is natural to consider variants of Definition 3.2.6 in which we replace the i.i.d. Rademacher random variables by a different choice of i.i.d. symmetric random variables. An especially important variant is obtained when using Gaussian random variables:

Definition 3.2.7 ((Empirical) Gaussian Complexities [60])**.** *Let* $\mathcal{Z}$ *be a data space, let* $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ *be an* $\mathbb{R}$*-valued function class, and let* $S = \{\boldsymbol{z}_i\}_{i=1}^m$ *be a data set consisting of* $m \in \mathbb{N}_{\geq 1}$ *data points* $\boldsymbol{z}_i \in \mathcal{Z}$*. The* empirical Gaussian complexity *of* $\mathcal{H}$ *with respect to* S *is defined as*

$$\hat{\mathcal{G}}_S(\mathcal{H}) \coloneqq \mathop{\mathbb{E}}_{\gamma_i\sim N(0,1)}\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \gamma_i h(\boldsymbol{z}_i)\right]\,, \tag{3.2.10}$$

where $\gamma_1, \ldots, \gamma_m$ *are i.i.d. standard Gaussian random variables.*
If $P \in \mathrm{Prob}(\mathcal{Z})$*, then the* Gaussian complexities *of* $\mathcal{H}$ *with respect to* P *are defined as*

$$\mathcal{G}_m(\mathcal{H}) \coloneqq \mathbb{E}_{S\sim P^m}\left[\hat{\mathcal{G}}_S(\mathcal{H})\right] \qquad \text{for } m \in \mathbb{N}_{\geq 1}\,, \tag{3.2.11}$$

where $S \sim P^m$ *means that* $S = \{\boldsymbol{z}_i\}_{i=1}^m$ *with the* $\boldsymbol{z}_i$ *drawn i.i.d. according to* P*.*

In fact, empirical Rademacher complexities and empirical Gaussian complexities are closely related (see, e.g. [61, Eqs. (4.8) and (4.9)]):

Theorem 3.2.8 (Empirical Rademacher Complexities versus Empirical Gaussian Complexities)**.** *There are universal constants* $c, C > 0$ *such that, for any* $\mathbb{R}$*-valued function class* $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ *and for any data set* $S = \{\boldsymbol{z}_i\}_{i=1}^m$ *consisting of* $m \in \mathbb{N}_{\geq 1}$ *data points* $\boldsymbol{z}_i \in \mathcal{Z}$*,*

$$c \cdot \hat{\mathcal{R}}_S(\mathcal{H}) \leq \hat{\mathcal{G}}_S(\mathcal{H}) \leq C \cdot \sqrt{\log(m)} \cdot \hat{\mathcal{R}}_S(\mathcal{H})\,. \tag{3.2.12}$$

Theorem 3.2.8 implies that for many purposes, we can treat empirical Rademacher complexities and empirical Gaussian complexities almost interchangeably.
The next theorem shows that we have data-dependent generalization guarantees in terms of empirical Rademacher complexities.

Theorem 3.2.9 (Generalization Bound via Empirical Rademacher Complexities)**.** *Let* $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ *and let* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$*. Write* $\mathcal{H} \coloneqq \{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \ell(y, f(\boldsymbol{x})) \mid f \in \mathcal{F}\}$*. For every*

$P \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ and for every $\delta \in (0,1)$, with probability $\geq 1-\delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,

$$\sup_{h \in \mathcal{F}} \left(R(h) - \hat{R}_S(h) \right) \leq 2 \cdot \hat{\mathcal{R}}_S(\mathcal{H}) + C \cdot \sqrt{\frac{\ln\left(1/\delta\right)}{m}}\,, \tag{3.2.13}$$

where $C > 0$ is a universal constant.

$$\mathbb{P}_{S \sim P^m}\left[\sup_{h \in \mathcal{F}} \left| R(h) - \hat{R}_S(h) \right| > \tfrac{C_1}{\sqrt{m}} \int_{\varepsilon}^{\sup_{h \in \mathcal{H}} \|h\|_{2,S}} \sqrt{\log \mathcal{N}(\mathcal{H}, \|\cdot\|_{2,S}, \beta)}\, \mathrm{d}\beta + C_2 \cdot \sqrt{\tfrac{\ln(1/\delta)}{m}} \right] \geq 1 - \delta\,,$$

Such Rademacher complexity-based generalization guarantees, which can be proved using the so-called "ghost sample" symmetrization technique, go back to [60, 62]. We have chosen to present a version of the result that can, e.g., be found in [34, Theorem 3.3].

We highlight two important consequences of Theorem 3.2.9 for binary classification and for regression over the reals.

Corollary 3.2.10. *In the setting of Theorem 3.2.9, if $\mathcal{Y} = \{-1, 1\}$ and $\ell(y_1, y_2) = 1 - \delta_{y_1, y_2}$ is the 0-1-loss, then we have: For every $P \in \mathrm{Prob}(\mathcal{X} \times \mathbb{R})$ and for every $\delta \in (0,1)$, with probability $\geq 1-\delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,*

$$\sup_{h \in \mathcal{F}} \left(R(h) - \hat{R}_S(h) \right) \leq \hat{\mathcal{R}}_{S|_{\mathcal{X}}}(\mathcal{F}) + C \cdot \sqrt{\frac{\ln\left(1/\delta\right)}{m}}\,, \tag{3.2.14}$$

where $C > 0$ is a universal constant and we defined $S|_{\mathcal{X}} \coloneqq \{\boldsymbol{x}_i\}_{i=1}^m$.

Corollary 3.2.11. *In the setting of Theorem 3.2.9, if $\mathcal{Y} = \mathbb{R}$ and ℓ is L-Lipschitz in the second argument (for any fixed first argument), then we have: For every $P \in \mathrm{Prob}(\mathcal{X} \times \mathbb{R})$ and for every $\delta \in (0,1)$, with probability $\geq 1-\delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,*

$$\sup_{h \in \mathcal{F}} \left(R(h) - \hat{R}_S(h) \right) \leq 2L \cdot \hat{\mathcal{R}}_{S|_{\mathcal{X}}}(\mathcal{F}) + C \cdot \sqrt{\frac{\ln\left(1/\delta\right)}{m}}\,, \tag{3.2.15}$$

where $C > 0$ is a universal constant and we defined $S|_{\mathcal{X}} \coloneqq \{\boldsymbol{x}_i\}_{i=1}^m$.

Corollary 3.2.10 can be obtained from Theorem 3.2.9 after observing that, for the output space $\mathcal{Y} = \{-1, 1\}$ and for the 0-1-loss ℓ, we have $\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{1}{2}\hat{\mathcal{R}}_{S|_{\mathcal{X}}}(\mathcal{F})$ for every data set S, see, e.g., [36, Section 1.8] for a proof. To obtain Corollary 3.2.11, we can apply Talagrand's Lemma (going back to [61]; see also [34, Lemma 5.7]).

We emphasize once more that, in contrast to Theorem 3.2.2, (empirical) Rademacher complexities lead to guarantees that depend on the specific training data set S or, using the strong concentration of empirical Rademacher complexities around their mean, on the underlying distribution P. A further difference between the guarantees of Theorem 3.2.2 and Theorem 3.2.9 is that the latter gives only one-sided bounds on the generalization error, whereas the former bounds the absolute generalization error. However, we can obtain also two-sided generalization

error bounds in terms of an expected supremum of a random process, if we consider not the empirical Rademacher complexities as in Definition 3.2.6, but a variant with an additional absolute value. That is, we can employ the data-dependent complexity measure

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim U(\{-1,1\}^m)}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{z}_i)\right|\right] \tag{3.2.16}$$

for a function class $\mathcal{H}\subseteq\mathbb{R}^{\mathcal{Z}}$, given a data set S of size m. Sometimes, also Eq. (3.2.16) is taken as the definition of "Rademacher complexity" in the literature.
To conclude our non-exhaustive overview over different complexity measures, we present a further data-dependent way of measuring the complexity of a function class. To this end, we first recall the notion of covering numbers in (pseudo-)metric spaces:

Definition 3.2.12 (Covering Numbers and Metric Entropies (see, e.g., [63, Definition 4.2.1]))**.** *Let (X,d) be a (pseudo-)metric space. Let $K\subseteq X$ and let $\varepsilon>0$. We call $N\subseteq X$ an* (interior) ε-covering net *of K if for all $x\in K$ there exists an $y\in N$ such that $d(x,y)\leq\varepsilon$.*
The (interior) ε-covering number $\mathcal{N}(K,d,\varepsilon)$ *is defined as the smallest possible cardinality of an ε-covering net of K. The* ε-metric entropy *is $\log_2\mathcal{N}(K,d,\varepsilon)$, the logarithm of the ε-covering number.*

For our purposes, covering numbers and metric entropies with respect to the pseudometric arising from the following seminorm will be particularly important:

Definition 3.2.13 (Empirical p-Seminorm)**.** *Let $\mathcal{Z}$ be some data space, let $p\in[1,\infty]$, and let $S=\{\boldsymbol{z}_i\}_{i=1}^m$ be a data set consisting of $m\in\mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i\in\mathcal{Z}$. We define the* empirical p-seminorm $\|\cdot\|_{p,S}$ *on $\mathbb{R}^{\mathcal{Z}}$ as*

$$\|h\|_{p,S}\coloneqq\left(\frac{1}{m}\sum_{i=1}^{m}|h(\boldsymbol{z}_i)|^p\right)^{\frac{1}{p}},\qquad \text{for } h\in\mathbb{R}^{\mathcal{Z}}. \tag{3.2.17}$$

The seminorm defined in Eq. (3.2.17) can be thought of as an L_p-norm when integrating against the probability measure given by a uniform distribution over the data set. In particular, this perspective explains the normalizing factor $1/m$. Because of this normalization, we see the following monotonicity behavior of these seminorms: For any data set S, if $1\leq p\leq q$, then $\|\cdot\|_{p,S}\leq\|\cdot\|_{q,S}$.
Now, by combining Definitions 3.2.12 and 3.2.13, we arrive at the following:

Definition 3.2.14 (Empirical Covering Numbers and Empirical Metric Entropies)**.** *Let $\mathcal{Z}$ be some data space, let $p\in[1,\infty]$, and let $S=\{\boldsymbol{z}_i\}_{i=1}^m$ be a data set consisting of $m\in\mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i\in\mathcal{Z}$. Let $\varepsilon>0$. The* ε-empirical covering number *of a function class $\mathcal{H}\subseteq\mathbb{R}^{\mathcal{Z}}$ with respect to S is defined as $\mathcal{N}(\mathcal{H},\|\cdot\|_{p,S},\varepsilon)$, the covering number for the set $\mathcal{H}$ in the seminormed space $(\mathbb{R}^{\mathcal{Z}},\|\cdot\|_{p,S})$. Accordingly, the* ε-empirical metric entropy *of $\mathcal{H}$ with respect to S is $\log_2\mathcal{N}(\mathcal{H},\|\cdot\|_{p,S},\varepsilon)$.*

Empirical covering numbers and their non-empirical counterparts, defined as covering numbers in the Hilbert space $L_p(P)$ of functions that are p-integrable against the probability measure P, come from the study of Banach spaces through random processes, see, e.g., [64, 65].

The monotonicity of empirical p-seminorms observed above implies a corresponding montonicity for empirical covering numbers and metric entropies: For any data set S, for any $\varepsilon > 0$, and for any function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$, if $1 \leq p \leq q$, then $\mathcal{N}(\mathcal{H}, \|\cdot\|_{p,S}, \varepsilon) \leq \mathcal{N}(\mathcal{H}, \|\cdot\|_{q,S}, \varepsilon)$.
Again, we can justifiably see empirical covering numbers as capturing the complexity of a function class because we can use them to derive PAC generalization guarantees:

Theorem 3.2.15 (Generalization Bound via Empirical Covering Numbers (see, e.g., [33, Theorem 16.5] and [36, Theorem 1.18])). *Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ and let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,1]$. Write again $\mathcal{H} \coloneqq \{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \ell(y, f(\boldsymbol{x})) \mid f \in \mathcal{F}\}$. For every $P \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ and for every $\varepsilon \in (0,1)$,*

$$\mathbb{P}_{S \sim P^m}\left[\sup_{h \in \mathcal{F}} \left|R(h) - \hat{R}_S(h)\right| < \varepsilon\right] \geq 1 - C_1 \cdot \left(\sup_{\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_{2m} \in \mathcal{Z}} \mathcal{N}(\mathcal{H}, \|\cdot\|_{1, \{\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_{2m}\}}, \varepsilon/C_2)\right) \cdot e^{-\frac{m\varepsilon^2}{C_3}}, \tag{3.2.18}$$

where $C_1, C_2, C_3 > 0$ are universal constants.

The core ingredients towards proving a result like Theorem 3.2.15 typically are ghost sample symmetrization and a union bound over a covering net. As we will comment on in more detail in Section 3.3, a different proof strategy leads to generalization guarantees that are usually slightly tighter, if we have control of empirical covering numbers with respect to the empirical 2-seminorm.

3.3 Relations Between Complexity Measures

A variety of relations between the different complexity measures introduced in Section 3.2 are known. In this section, we collect some of them. We begin with different ways of upper bounding empirical covering numbers and metric entropies in terms of other complexity measures. First, we recall how the different combinatorial complexity measures of VC dimension, pseudo-dimension, and fat-shattering dimension can be used to upper bound empirical covering numbers. We start with a result for the VC dimension:

Theorem 3.3.1 (Empirical Covering Numbers versus VC Dimension (see, e.g., [63, Theorem 8.3.18 and its proof])). *There is a universal constant $C > 0$ such that , for any $\{0,1\}$-valued function class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{Z}}$, for any data set $S = \{\boldsymbol{z}_i\}_{i=1}^m$ consisting of $m \in \mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i \in \mathcal{Z}$, and for any $\varepsilon \in (0,1)$,*

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{2,S}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{C \cdot \mathrm{VCdim}(\mathcal{H})}. \tag{3.3.1}$$

Thus, the VC dimension gives rise to covering number bounds with respect to the empirical 2-seminorm that look very much like standard covering number bounds for norm balls in Euclidean space in terms of the dimension of the space. Next, we present covering number bounds in terms of pseudo- and fat-shattering dimension:

Theorem 3.3.2 (Empirical Covering Numbers versus Pseudo- and Fat-Shattering Dimension ([66, Theorem 1], see also [33, Sections 12 and 18] and [67, Sections 4.2.2 and 4.2.4])). *Let $p \in [1, \infty)$. There are universal constants $c, C > 0$ such that , for any $[0,1]$-valued function class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{Z}}$, for any data set $S = \{\boldsymbol{z}_i\}_{i=1}^m$ consisting of $m \in \mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i \in \mathcal{Z}$, and for any $\varepsilon \in (0,1)$,*

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{p,S}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{C \cdot \mathrm{fat}(\mathcal{H}, c\varepsilon)} \leq \left(\frac{2}{\varepsilon}\right)^{C \cdot \mathrm{Pdim}(\mathcal{H})} . \tag{3.3.2}$$

Moreover, there exist universal constants $\tilde{C}, \tilde{c} > 0$ such that, if $m \geq \mathrm{fat}(\mathcal{H}, \tilde{c}\varepsilon)$, then

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{p,S}, \varepsilon) \geq 2^{\tilde{C} \cdot \mathrm{fat}(\mathcal{H}, \tilde{c}\varepsilon)} . \tag{3.3.3}$$

Thus, upper bounds on combinatorial dimensions imply empirical covering number upper bounds with respect to any p-seminorm. Note that, in particular, for the case of a $\{0,1\}$-valued function class, pseudo-dimension and VC dimension are equal, so Theorem 3.3.2 leads to a version of Theorem 3.3.1 for general $p \in [1, \infty)$. Interestingly, due to the second part of Theorem 3.3.2, we also have a converse in the case of the fat-shattering dimension: Empirical covering number upper bounds imply upper bounds on the fat-shattering dimension. Such a converse, however, is not possible for the pseudo-dimension, as discussed, e.g., in [33, Section 12.5].

Empirical covering numbers are also closely related to Gaussian and Rademacher complexities. Namely, we can upper bound empirical Rademacher complexities in terms of an integral over square roots of empirical metric entropies:

Theorem 3.3.3 (Dudley's Theorem [68]). *For a fixed data set $S = \{\boldsymbol{z}_i\}_{i=1}^m$ consisting of $m \in \mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i \in \mathcal{Z}$, let $\mathcal{H}$ be a subset of the pseudo-metric space $(\mathbb{R}^{\mathcal{Z}}, \|\cdot\|_{2,S})$ and let $\gamma_0 \coloneqq \sup_{h \in \mathcal{H}} \|h\|_{2,S}$. Then the empirical Rademacher complexity $\hat{\mathcal{R}}_S(\mathcal{H})$ of $\mathcal{H}$ with respect to S can be upper bounded as*

$$\hat{\mathcal{R}}_S(\mathcal{H}) \leq \inf_{\varepsilon \in [0, \gamma_0/2)} \left\{ 4\varepsilon + \frac{12}{\sqrt{m}} \int_{\varepsilon}^{\gamma_0} \sqrt{\log \mathcal{N}(\mathcal{H}, \|\cdot\|_{2,S}, \beta)} \, \mathrm{d}\beta \right\} . \tag{3.3.4}$$

We have stated a version of Dudley's Theorem that can, e.g., be found in [36, Theorem 1.19]. As discussed in [63, Remark 8.1.5] a similar result also holds for the variant of the Rademacher complexity that involves an additional absolute value, as introduced in Eq. (3.2.16), assuming that the class $\mathcal{H}$ contains the zero function.

Theorem 3.3.3 can be combined with Theorem 3.2.9 to obtain yet another generalization bound in terms of empirical covering numbers. The generalization guarantee obtained in this way is formulated via covering numbers for the empirical 2-seminorm. This is in contrast to Theorem 3.2.15, which considered covering with respect to the empirical 1-seminorm. If, however, we can control the empirical covering numbers $\mathcal{N}(\mathcal{H}, \|\cdot\|_{2,S}, \varepsilon)$, using the combination of Theorems 3.2.9 and 3.3.3 – or, to be more precise, their respective versions with a reinstated absolute value – can lead to a slightly tighter generalization bound than the one obtained by plugging the bound on $\mathcal{N}(\mathcal{H}, \|\cdot\|_{2,S}, \varepsilon)$ into Theorem 3.2.15.

Moreover, we can lower bound empirical Gaussian complexities in terms of square roots of empirical metric entropies. This follows from a more general result about mean-zero Gaussian processes, Sudakov's minoration inequality, which can, e.g., be found in [63, Theorem 7.4.1]. We state it here only for the special case relevant to our learning-theoretic framework:

Theorem 3.3.4 (Sudakov's Minoration Inequality for Empirical Gaussian Complexities)**.** *For a fixed data set $S = \{\boldsymbol{z}_i\}_{i=1}^m$ consisting of $m \in \mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i \in \mathcal{Z}$, let $\mathcal{H}$ be a subset of the pseudo-metric space $(\mathbb{R}^{\mathcal{Z}}, \|\cdot\|_{2,S})$. Then the empirical Gaussian complexity $\hat{\mathcal{G}}_S(\mathcal{H})$ of $\mathcal{H}$ with respect to S can be lower bounded as*

$$\hat{\mathcal{G}}_S(\mathcal{H}) \geq \frac{C}{\sqrt{m}} \cdot \sup_{\varepsilon \geq 0} \left\{ \varepsilon \cdot \sqrt{\log \mathcal{N}(\mathcal{H}, \|\cdot\|_{2,S}, \varepsilon)} \right\}, \tag{3.3.5}$$

where $C > 0$ is a universal constant.

According to Theorem 3.2.8, Theorem 3.3.4 also allows us to lower bound empirical Rademacher complexities via empirical covering numbers. The scaling with training data size m in the obtained lower bound matches that in Dudley's upper bound of Eq. (3.3.4) up to a logarithmic factor.
The above relations can now be combined to establish further connections between complexity measures, such as the following:

Corollary 3.3.5 (Rademacher Complexities versus VC Dimension)**.** *There is a universal constant $C > 0$ such that, for any $\{0,1\}$-valued function class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{Z}}$ and for any data set data set $S = \{\boldsymbol{z}_i\}_{i=1}^m$ consisting of $m \in \mathbb{N}_{\geq 1}$ data points $\boldsymbol{z}_i \in \mathcal{Z}$,*

$$\hat{\mathcal{R}}_S(\mathcal{H}) \leq C \cdot \sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{m}}. \tag{3.3.6}$$

We highlight this last connection because, together with Theorem 3.2.9, it can be used to prove Theorem 3.2.2. (Again, to get two-sided generalization error bounds, we can reinstate absolute values in Theorem 3.2.9 and Corollary 3.3.5.)
There are several further approaches towards generalization bounds beyond the one based on complexity measures. In particular, whereas the perspective of complexity measures focuses on the hypothesis classes used by the learner, other approaches take additional properties of the learning procedure into account. Important examples include sample compression schemes [69], PAC-Bayesian generalization bounds [41], algorithmic stability [70], or differential privacy [71]. Discussing these topics is, unfortunately, beyond the scope of this thesis.

3.4 Undecidable Problems in Classical Learning Theory

So far in this chapter, we have presented the PAC framework as an approach to mathematically formalize statistical questions in machine learning. And we have seen ways of obtaining theoretical learnability guarantees, in the form of generalization bounds, from different complexity measures. While this shows the usefulness of these complexity measures, we have mostly ignored the question of how to evaluate them. For example, in the setting of PAC binary classification, we

have seen that the VC dimension determines whether a class is learnable, and even characterizes the corresponding sample complexity. However, we have not described a general procedure for evaluating the VC dimension of a hypothesis class. In this section, we present a high-level discussion regarding some subtleties in the approach of determining learnability through complexity measures, from the perspective of formal logic and computability theory.

Prior work has emphasized different aspects of logic and computability in connection to learning theory. Here, we highlight only some of these directions and refer to [7, Section 1.2] for a more extensive discussion of prior work. For instance, [72] investigated the learnability of uncomputable problems. Also, [73] considered the computational complexity of deciding finiteness of the VC dimension, in particular showing this to be an undecidable problem. Later, this was reproved and interpreted from a philosophical perspective in the context of the problem of induction by [74]. Recently, [75] demonstrated that learnability can in general be undecidable in the sense of formal logic. Motivated by this development, [76] proposed a variant of PAC learning in which learners have to be computable functions and investigated connections between standard PAC learnability and computable PAC learnability in binary classification problems. This (incomplete) list of references already demonstrates that research on (un-)decidability in learning theory is multi-faceted and, despite a history of over 20 years, still ongoing. In the remainder of this section, we describe how our work [7] adds to this research effort.

Section 3.2 introduced complexity measures as a tool for answering the fundamental question of when (PAC) learning is possible. This question, however, is potentially different from the question of whether we can *decide* when (PAC) learning is possible. Concretely, we can understand the verb "decide" in the previous sentence to mean "mathematically prove" or "algorithmically determine". For these two interpretations, we know from Gödel's Incompleteness Theorems [77] and Turing's Halting Problem [78] that we cannot always succeed at "deciding." This now raises the following questions: First, if a hypothesis class is learnable, can we prove it to be so (in a formal system of interest)? And second, is there a universal algorithmic procedure for determining whether a hypothesis class is learnable?

According to the results of Section 3.2, if we focus on the learning framework of PAC binary classification with respect to the 0-1-loss, PAC learnability becomes equivalent to finiteness of the VC dimension, a purely combinatorial property of the respective hypothesis class. Thus, in this setting, we can translate the two questions above as: First, if a hypothesis class has finite VC dimension, can we prove this finiteness? And second, is there a universal algorithmic procedure for determining whether a hypothesis class has finite VC dimension? In [7], we find the following two negative answers to these questions:

Theorem 3.4.1 (Informal Version of [7, Corollary 2.11])**.** *There are hypothesis classes with finite VC dimension for which we cannot prove that the VC dimension is finite.*

Theorem 3.4.2 (Informal Version of [7, Corollary 2.20])**.** *There is no general-purpose algorithm for deciding whether a hypothesis class has a finite VC dimension.*

Taken together, Theorems 3.4.1 and 3.4.2 tell us that, in general, we can neither prove nor algorithmically determine finiteness of the VC dimension. Thus, finiteness of the VC dimension, and with it PAC learnability in binary classification problems are undecidable. This result carries

over to PAC binary classification using quantum examples, since also there the VC dimension characterizes learnability, compare [79]. This undecidability is not specific to the VC dimension. In fact, as we argue in [7], combinatorial complexity parameters relevant to teaching problems and to online learning settings share this property. So, this undecidability also translates to learnability in teacher-learner interactions [46] and to different notions of online learnability [44, 80]. In this sense, we have demonstrated in [7] that certain basic questions in different learning frameworks are undecidable.

We conclude this section with a short discussion of how our results in [7] relate to two of the prior works mentioned above. First, the undecidability of finiteness of the VC dimension stated informally in Theorem 3.4.2 can be obtained as a Corollary of [73, Theorem 4.1] and of [74, Theorem 1]. Both [73] and [74] prove a stronger result, namely the so-called Σ_2-completeness of deciding finiteness of the VC dimension, deriving it as a consequence of the Σ_2-completeness of deciding finiteness of the domain of a partial recursive function. Compared to this line of reasoning, our proof strategy in [7] has mainly two advantages. On the one hand, we give a construction that allows to prove undecidability not only for the VC dimension, but also for other combinatorial complexity measures, and not only in the sense of uncomputability, but also in the sense of independence of the axioms in a formal system. On the other hand, we use no results from formal logic and computability theory beyond the arguably most fundamental undecidable problems, Gödel's second incompleteness theorem and Turing's halting problem.

Second, while also [75] proved the logical undecidability of a certain learning problem, tracing it back to the independence of the continuum hypothesis from the axioms of Zermelo-Fraenkel set theory (including the axiom of choice), their scenario differs from our setting in [7] in at least two important ways. While [75] show that learnability for their learning problem cannot be characterized by a dimension-like parameter akin to the VC dimension, our undecidability results in [7] are for learning problems that admit such a characterization. In fact, we make use of exactly such dimension-like parameters (the VC, teaching, and Littlestone dimensions) in our proofs. And whereas the results in [75] rely on the use of the continuum, we formulate our results in [7] using only natural numbers and computable constructions. In particular, this allows us to also obtain uncomputability results and, relying on insights from [76], to immediately extend our results from PAC binary classification also to computable PAC binary classification.

To conclude this chapter: The complexity measures of Section 3.2 are powerful mathematical tools for understanding generalization bounds for PAC learning. Still, there are further computational questions in learning theory beyond the probability theory and statistics of generalization. Our deliberations on undecidability outline merely one of many possibilities of emphasizing algorithmic and computational aspects of learning theory. In the next two chapters, we now take a quantum computing perspective on learning, and also a learning perspective on quantum information.

Chapter 4

Variational Quantum Machine Learning

This chapter discusses a field lying at the intersection point of these two theories, namely the theory of variational quantum machine learning (QML). More precisely, we focus on questions of generalization in variational QML: We discuss generalization error bounds for supervised machine learning models based on parametrized quantum circuits (PQCs) that are trained via classical optimization.
The structure of this chapter is as follows: In Section 4.1, we describe how to use PQCs for machine learning on either classical or quantum data. Having defined the machine learning models used in variational QML, we then proceed by giving an overview over generalization bounds for such models in Section 4.2. In this discussion, we in particular distinguish between generalization guarantees arising from properties of the trainable part of the PQC (Subsection 4.2.1) and, when using variational QML for classical data, generalization guarantees derived from the classical-to-quantum data-encoding (Subsection 4.2.2).

4.1 Parametrized Quantum Circuits for Machine Learning

Variational Quantum Algorithms: In recent years, *variational quantum algorithms (VQAs)* have established themselves as a promising candidate for applications of near-term quantum computing architectures. VQAs are based on *parametrized quantum circuits (PQCs)*, i.e., quantum circuits containing gates that depend on classical parameters. Crucially, while the quantum circuits themselves are run on a quantum computer, the classical parameters are trained using classical optimization with respect to a certain target function. More precisely, the quantum computer is typically used in the evaluation of the target function and its gradient, the latter often achieved through so-called parameter-shift rules [81, 82], but then the remaining parameter optimization is performed classically. This delegation of the optimization task to a classical computer is what can make VQAs viable already on noisy intermediate-scale quantum devices, in what was termed the "NISQ era" by John Preskill [83]. Among the early works on VQAs, in particular the variational quantum eigensolver (VQE) [84] for approximating ground state energies of a Hamiltonian and the Quantum Approximate Optimization Algorithm (QAOA) for approximating the solutions of combinatorial optimization problems [85] have spurred further research into the potential of such hybrid quantum-classical methods. In this chapter, we focus

specifically on the use of PQCs for machine learning. We refer interested reader to the review [86] for a broader overview over VQAs.

PQCs and Variational QML: As described above, PQCs lead to classically trainable models and therefore naturally lend themselves to applications in machine learning by optimizing the parameters in the circuit to perform well on a given training data set. We will refer to QML models implemented by classically optimizing a PQC on data as *PQC-based QML models* or *variational QML models*. Since the first proposals of variational QML in the influential works [87–89], this field has attracted significant interest. In particular, works such as [90–94] are beginning to understand challenges arising when training variational QML models. Moreover, several works present numerical experiments exploring applications of variational QML, as for example [87, 88, 95–97]. For the purposes of this thesis, however, we focus on a mathematically rigorous analysis of generalization in variational QML.

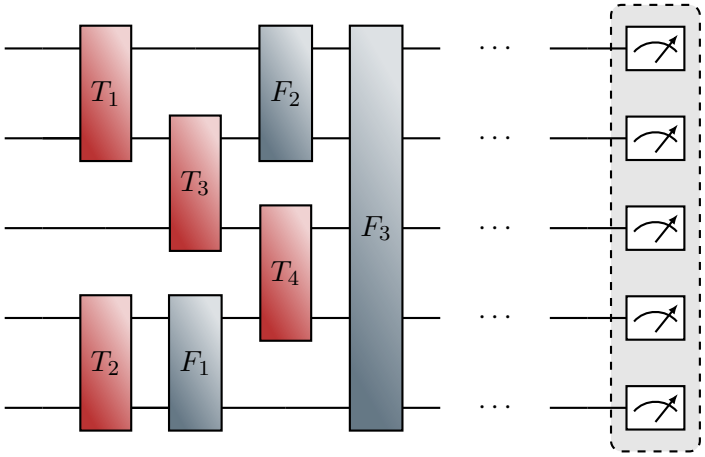


Figure 4.1: A schematic of a PQC for quantum data, here for $n = 5$ qudits and locality parameter $k = 2$. The circuit, which can act on any n-qudit input state, consists of k-local trainable gates $T_i = T_i(\boldsymbol{\theta})$ (red) and fixed (potentially global) gates F_ℓ (blue). At the output of a circuit, a (potentially global) measurement is performed.

Mathematical Framework for PQC-based QML Models: We dedicate the remainder of this subsection to a careful formulation of a mathematical framework formalizing the discussion above. We begin by presenting PQCs that can be used to learn from quantum data. We describe such a PQC by a quantum circuit on n qudits, schematically depicted in Fig. 4.1, consisting of two types of gates: On the one hand, it contains trainable gates T_i, which we assume to be k-local for some n-independent $k \leq n$. These gates are trainable in the sense that they depend on some parameter vector $\boldsymbol{\theta}$ with real entries, that is, $T_i = T_i(\boldsymbol{\theta})$. Note: We do not require the trainable gates to be geometrically local. For example, for $k = 2$ we also allow trainable gates to act on two non-neighboring qudits. On the other hand, the circuit contains fixed gates F_ℓ, which are allowed to be global. Importantly, these fixed gates are not trainable. In this thesis, quantum circuits and the gates therein are not restricted to unitaries but can be CPTP maps. Thus, we ascribe the variational QML model corresponding to such a PQC on n qubits the parametrized

CPTP map $T_{\boldsymbol{\theta}}^{\mathrm{QMLM}} \in \mathcal{CPTP}_{d^n}$ obtained by composing the circuit gates according to the circuit structure.

Viewed in this way, a variational QML model naturally acts on quantum data, since quantum states are the natural inputs for CPTP maps. When aiming to use a PQC-based QML model for learning with classical data, we can add a further type of gate to the PQC. Namely, as depicted in Fig. 4.2, we may add an initial circuit layer with a (potentially global) CPTP map E that depends on the classical data input $\boldsymbol{x}$, that is $E = E(\boldsymbol{x})$. If we fix the input state to be $(|0\rangle\langle 0|)^{\otimes n}$, the first layer implements a classical-to-quantum data encoding $\boldsymbol{x} \mapsto \rho(\boldsymbol{x}) \coloneqq E(\boldsymbol{x})\,((|0\rangle\langle 0|)^{\otimes n})$, on which the remaining PQC then acts. We will refer to such a PQC as *encoding-first* PQC. Mathematically, we can describe an encoding-first PQC either by considering the encoding $\boldsymbol{x} \mapsto \rho(\boldsymbol{x})$ and the parametrized CPTP map $T_{\boldsymbol{\theta}}^{\mathrm{QMLM}} \in \mathcal{CPTP}_{d^n}$ separately, or by viewing the whole PQC as implementing a CPTP map $T_{\boldsymbol{x},\boldsymbol{\theta}}^{\mathrm{QMLM}} \in \mathcal{CPTP}_{d^n}$, depending on both the inputs $\boldsymbol{x}$ and the choice of parameters $\boldsymbol{\theta}$, where we focus on the action of $T_{\boldsymbol{x},\boldsymbol{\theta}}^{\mathrm{QMLM}}$ on $(|0\rangle\langle 0|)^{\otimes n}$. Notice that, with this choice of notation, $T_{\boldsymbol{x},\boldsymbol{\theta}}^{\mathrm{QMLM}}((|0\rangle\langle 0|)^{\otimes n}) = T_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\rho(\boldsymbol{x}))$.

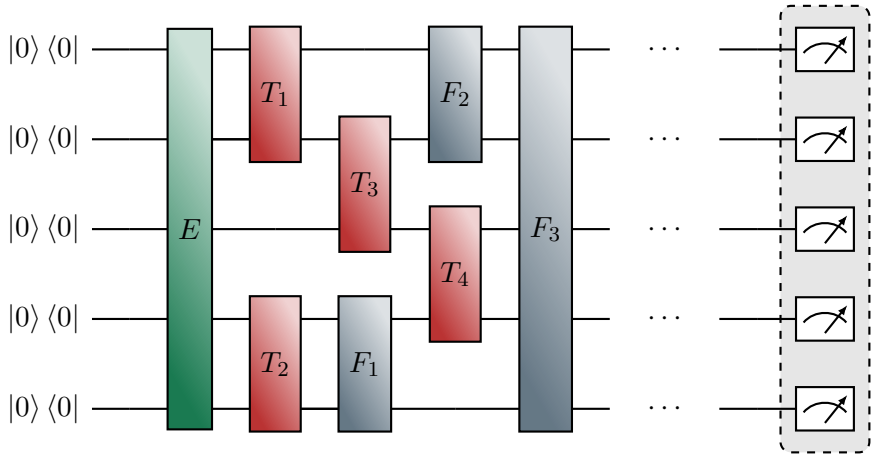


Figure 4.2: A schematic of an encoding-first PQC for classical data, here for $n = 5$ qudits and locality parameter $k = 2$. The circuit, acting on the $(|0\rangle\langle 0|)^{\otimes n}$-state, consists of an encoding gate $E = E(\boldsymbol{x})$ (green), preparing the data-encoding state $\rho(\boldsymbol{x})$, followed by k-local trainable gates $T_i = T_i(\boldsymbol{\theta})$ (red) and fixed (potentially global) gates F_ℓ (blue). At the output of a circuit, a (potentially global) measurement is performed.

While encoding-first PQCs already provide a way of incorporating classical data inputs, this is not the most general way. Namely, in the spirit of data re-uploading [98], we can allow for encoding gates to be placed throughout the circuit, as illustrated in Fig. 4.3. Such a *data re-uploading PQC* is then naturally described by a CPTP map $T_{\boldsymbol{x},\boldsymbol{\theta}}^{\mathrm{QMLM}} \in \mathcal{CPTP}_{d^n}$ that is "parametrized" both by data inputs $\boldsymbol{x}$ and trainable parameters $\boldsymbol{\theta}$. Again, we focus specifically on the action of $T_{\boldsymbol{x},\boldsymbol{\theta}}^{\mathrm{QMLM}}$ on the state $(|0\rangle\langle 0|)^{\otimes n}$. Note that encoding-first PQCs are a special case of data re-uploading PQCs, in which the encoding and the trainable parts of the PQC are separate from each other.

From now on, we will abuse notation in the interest of a unified presentation for the cases of quantum and classical data: We will write $T_{\boldsymbol{x},\boldsymbol{\theta}}^{\mathrm{QMLM}}$ for the CPTP map associated to a variational QML model, even if the QML model acts on quantum data. In this case, that is, for a quantum

input $\boldsymbol{x} = \rho \in \mathcal{X} = \mathcal{S}(\mathbb{C}^{d^n})$, we use this notation for a PQC acting on quantum data by formally identifying it with our notation for an encoding-first PQC in which we take the "encoding" $\mathcal{X} \ni \boldsymbol{x} \to \rho(\boldsymbol{x})$ as given by the identity map $\mathcal{X} \ni \boldsymbol{x} = \rho \mapsto \rho(\boldsymbol{x}) = \rho$. Continuing this abuse of notation, in this case we also write $T^{\mathrm{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}((|0\rangle\langle 0|)^{\otimes n}) = T^{\mathrm{QMLM}}_{\boldsymbol{\theta}}(\rho(\boldsymbol{x}))$ to mean $T^{\mathrm{QMLM}}_{\boldsymbol{\theta}}(\rho)$.

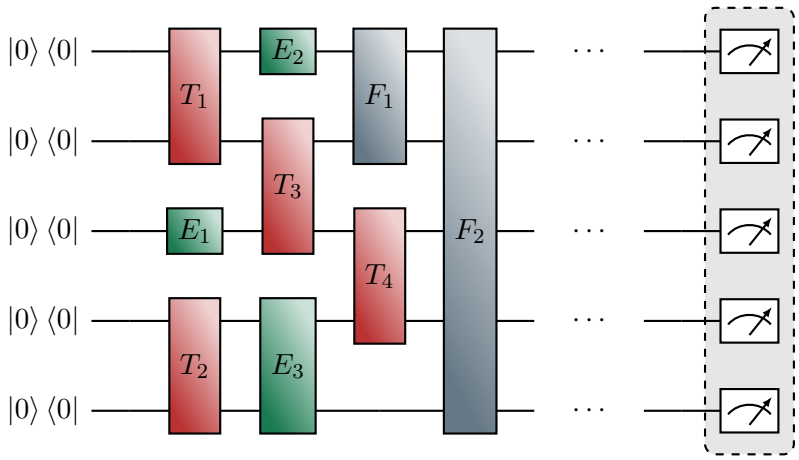


Figure 4.3: A schematic of a data re-uploading PQC for classical data, here for $n = 5$ qudits and locality parameter $k = 2$. The circuit, acting on the $(|0\rangle\langle 0|)^{\otimes n}$-state, consists of k-local trainable gates $T_i = T_i(\boldsymbol{\theta})$ (red), encoding gates $E_j = E_j(\boldsymbol{x})$ (green), and fixed (potentially global) gates F_ℓ (blue). At the output of a circuit, a (potentially global) measurement is performed.

Evaluating the Performance in Variational QML: To conclude our presentation of the mathematical setting in which we prove our generalization bounds, we discuss how to evaluate the performance of a PQC-based QML model. Here, we consider two options. For the first of the two, we fix a Hermitian observable $M \in \mathcal{B}(\mathbb{C}^{d^n})$. Then, given a variational QML model for processing classical data, with associated parametrized CPTP map $T^{\mathrm{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}} \in \mathcal{CPTP}_{d^n}$, for each choice of parameters $\boldsymbol{\theta}$, we can consider the function

$$f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}, \; f_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq \mathrm{tr}\left[M \cdot T^{\mathrm{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}\left((|0\rangle\langle 0|)^{\otimes n}\right)\right] . \tag{4.1.1}$$

That is, the function value $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ upon input $\boldsymbol{x}$ is the expectation value of the observable M when measured in the state $T^{\mathrm{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}((|0\rangle\langle 0|)^{\otimes n})$. Accordingly, the variational QML model implements the hypothesis class

$$\mathcal{F}^{\mathrm{QMLM}} \coloneqq \{f_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta\} \subseteq \mathbb{R}^{\mathcal{X}} , \tag{4.1.2}$$

where Θ is the set of admissible parameters. Now, we have arrived at a real-valued concept class for the QML model and can thus employ any loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ to evaluate the performance of a hypothesis from $\mathcal{F}^{\mathrm{QMLM}}$ and work in the PAC setting of Chapter 3.

For the second option, we use observables not to define a hypothesis class associated to the model, but directly to define a physically motivated notion of loss. Namely, for each (classical or quantum) input $\boldsymbol{x}$ and for each (classical or quantum) output y, we let $O^{\mathrm{loss}}_{\boldsymbol{x},y} \in \mathcal{B}(\mathbb{C}^{d^n})$ be

a Hermitian loss observable. Then, on the data point $(\boldsymbol{x}, y)$, the variational QML model with parameter setting $\boldsymbol{\theta}$ incurs the loss

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) \coloneqq \operatorname{tr}\left[O^{\text{loss}}_{\boldsymbol{x},y} \cdot T^{\text{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}((|0\rangle\langle 0|)^{\otimes n})\right] . \tag{4.1.3}$$

With this definition of loss in place, even without an intermediate hypothesis class, we can again work in the PAC framework of Chapter 3: Our goal is to achieve a small expected loss (with high probability), where the expectation is over a data-generating probability measure $P \in \text{Prob}(\mathcal{X} \times \mathcal{Y})$. And we study generalization guarantees to justify an approach towards this goal based on training the model to achieve small empirical risk on training data.

Variational QML and Kernel Methods: Above, we have introduced variational QML based on PQCs. At this point, we shortly mention a second influential perspective on quantum circuits in quantum machine learning, through the lens of kernel models. Kernel methods are a well established framework in classical machine learning theory [99]. Here, we embed training data into a (typically high- or even infinite-dimensional) feature space and then optimize a loss function over linear models in feature space. This optimization is analytically tractable for several cases of interest, for example using kernel ridge regression for a least squares loss with Tikhonov regularization (compare, e.g., [36, Section 3.6]). Importantly, combining the so-called "kernel trick" with the Representer Theorem [100, 101], solving this optimization and evaluating the obtained function does not require explicit knowledge of the feature map or computations in the high-dimensional feature space, but can be achieved given only the ability to compute inner products with the feature vectors associated to the training data instances. This makes kernel methods attractive both for theoretical analysis and numerical implementations.

In this spirit, one can study quantum kernel methods, that is, quantum machine learning using kernels that can be evaluated on a quantum computer, as advocated for, e.g., in [87]. Here, one considers mapping classical data to n-qudit quantum states, thereby embedding into the $d^n \times d^n$-dimensional feature space of Hermitian operators. In this feature space, linear models correspond to Hermitian observables. A detailed review of the literature on quantum kernels is beyond the scope of this thesis, we only discuss their connection to our framework of variational quantum machine learning. Quantum kernel methods are effectively equivalent to encoding-first PQC-based QML models in which we allow for PQCs of arbitrarily large depth [102, 103]. As pointed out by [102, 104], this means that quantum kernel methods outperform variational QML models with respect to training performance, however, potentially at the cost of a worse generalization performance. In the next section, we discuss how limiting the size and depth of PQCs used for QML leads to generalization guarantees. In particular, we do not work in the infinite-depth regime, which would give rise to quantum kernel methods. Therefore, our results hint at the "hidden" infinite depth in quantum kernel methods as a potential explanation for their sometimes problematic generalization behavior.

4.2 Generalization Guarantees for Variational Quantum Machine Learning

From the very definition of a PQC, three natural sources for "complexity" – thought of in the learning-theoretic sense of Section 3.2 – in a PQC-based QML model arise. First, the trainable elements in a PQC crucially influence the generalization performance of the corresponding variational QML model. This facet of generalization in variational QML is the focus of Subsection 4.2.1. Second, also the choice of the measurement performed at the output of the PQC can be important for generalization. In this thesis, we do not investigate this aspect of the problem in detail, but instead refer the interested reader to [105]. And third, when using a PQC-based QML model to process classical data, also the strategy for encoding the classical inputs into the quantum circuit is relevant from the perspective of generalization. We discuss this in Subsection 4.2.2.

While these three approaches to generalization in variational QML cover a significant fraction of the literature on rigorous generalization guarantees for PQC-based QML, some recent works have explored different paths. Before discussing generalization bounds derived from the trainable part and the encoding strategy in a PQC in more detail, we highlight some of these alternative approaches. On the one hand, [96] suggested to use the so-called empirical Fisher information matrix to define a new complexity measure, which they termed *effective dimension*, a local variant of which was recently considered for generalization in classical machine learning [106]. On the other hand, [107] proposed a quantum information-theoretic approach towards bounding both the generalization and the approximation error that a PQC-based model achieves in a classification task. Both of these works take a holistic perspective on the PQC, not separating trainable gates from the encoding gates or from the measurement. While this perspective has the advantage of incorporating the interplay between the different architectural elements in a PQC, and potentially even the training procedure, the resulting generalization error bounds are not easy to evaluate analytically except in simple examples. This is in contrast to the generalization bounds presented in this chapter.

4.2.1 Generalization Guarantees Based on the Trainable Part

When trying to bound the generalization error of machine learning models, classical or quantum, described by a concrete architecture, a natural approach is as follows: We bound the complexity of the corresponding function class, measured by one of the complexity measures introduced in Section 3.2, in terms of properties of the trainable elements in the model architecture. As a concrete classical example, we can aim to bound the complexity of a class of functions implemented by a classical neural network in terms of the number of adjustable parameters (weights and biases) in the network (see, e.g., [108] and [33, Chapter 14]). In this subsection, we present two ways of implementing this strategy of proving generalization bounds in the case of variational QML models based on PQCs with local trainable gates.

First, in the case of a variational QML model for processing classical data, we can bound the pseudo-dimension of the associated function class $\mathcal{F}^{\mathrm{QMLM}}$, introduced in our first variant of evaluating the performance of a QML model, in terms of the number of local trainable gates.

With the notation and framework as formulated in Section 4.1, we can now reformulate the main result of [1] as follows:

Theorem 4.2.1 (Pseudo-Dimension Bounds for variational QML (see [1, Theorems 2 and 4, Remark 1])). *Let $\mathcal{X} \coloneqq \{\boldsymbol{x} \in (\mathbb{C}^d)^{\otimes n} \mid \|\boldsymbol{x}\|_2 = 1\}$ for some $n, d \in \mathbb{N}_{\geq 1}$. Consider an encoding-first n-qudit PQC with $\Gamma \in \mathbb{N}$ 2-local trainable gates and the encoding map $E : \mathcal{X} \to \mathcal{CPTP}_{d^n}$, $E(\boldsymbol{x})(\rho) = \mathrm{tr}[\rho] \, |\boldsymbol{x}\rangle \langle \boldsymbol{x}|$, which leads to the classical-to-quantum data encoding $\boldsymbol{x} \mapsto \mathrm{tr}[(|0\rangle \langle 0|)^{\otimes n}] \, |\boldsymbol{x}\rangle \langle \boldsymbol{x}| = |\boldsymbol{x}\rangle \langle \boldsymbol{x}|$. Fix the Hermitian observable $M = (|0\rangle \langle 0|)^{\otimes n} \in \mathcal{B}(\mathbb{C}^{d^n})$. Then, the pseudo-dimension of the associated function class $\mathcal{F}^{\mathrm{QMLM}} \subseteq [0,1]^{\mathcal{X}}$ satisfies*

$$\mathrm{Pdim}(\mathcal{F}^{\mathrm{QMLM}}) \leq C \cdot \Gamma \log_2(\Gamma) \,, \tag{4.2.1}$$

where $C = C(d) > 0$ depends polynomially on d.

The proof of Theorem 4.2.1 is inspired by [109], which bounded the VC dimension of semi-algebraic function classes in terms of the number and degrees of the involved polynomials. First, we show that hypotheses in $\mathcal{F}^{\mathrm{QMLM}}$ can be written as polynomials with real coefficients, with the entries of the trainable gates and the entries of the input vector as variables. Importantly, we prove that the degree of these polynomials in the variables associated to the trainable gates is bounded by $2d^8 \cdot \Gamma$. This allows us to reduce questions of pseudo-shattering (as defined in Definition 3.2.4) to questions of consistent sign assignments to polynomials. Similarly to [109], we can now employ a result due to [9], which bounds the number of such sign assignments in terms of the number and degrees of the involved polynomials, to obtain an upper bound on the maximal size of a pseudo-shattered set and thus on the pseudo-dimension of $\mathcal{F}^{\mathrm{QMLM}}$.
Theorem 4.2.1 shows that, assuming a simple classical-to-quantum encoding strategy and a specific observable, the learning-theoretic complexity of a variational QML model, as measured by the pseudo-dimension of the associated function class, grows at worst slightly superlinearly with the number of trainable gates in the PQC. It turns out that both the restriction on the data-encoding and on the observable can be lifted and such a statement remains valid. In [6], we prove a corresponding complexity bound for general PQCs, this time in terms of metric entropies and for our second way of evaluating the performance in variational QML.

Theorem 4.2.2 (Metric Entropy Bounds for variational QML Models [6, Theorem 7, Proof of Theorem 11]). *Let $\mathcal{X}, \mathcal{Y}$ be some (classical or quantum) input and output spaces, respectively. Consider an n-qudit PQC with $\Gamma \in \mathbb{N}$ k-local trainable gates. Fix Hermitian loss observables $O^{\mathrm{loss}}_{\boldsymbol{x},y} \in \mathcal{B}(\mathbb{C}^{d^n})$ with the corresponding loss function as in Eq. (4.1.3). Then, for any training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, the empirical metric entropies of the function class*

$$\mathcal{G}^{\mathrm{QMLM}} \coloneqq \{\mathcal{X} \times \mathcal{Y} \ni (\boldsymbol{x}, y) \mapsto \ell(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{\theta} \in \Theta\} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} \tag{4.2.2}$$

with respect to S satisfy

$$\log_2 \mathcal{N}(\mathcal{G}^{\mathrm{QMLM}}, \|\cdot\|_{\infty, S|_{\mathcal{X}}}, \varepsilon) \leq C \cdot \Gamma \log_2 \left(\frac{C_{\mathrm{loss}} \cdot \Gamma}{\varepsilon} \right) , \tag{4.2.3}$$

for any $\varepsilon \in (0, 1]$*, where* $C = C(d, k) > 0$ *depends polynomially on* d *and exponentially on* k*, and we defined the quantity* $C_{\text{loss}} \coloneqq \sup_{\boldsymbol{x},y} \left\| O^{\text{loss}}_{\boldsymbol{x},y} \right\|$.

The two crucial steps in proving Theorem 4.2.2 are: First, regard the space of possible choices for a single k-qudit gate as a compact subset of a complex vector space whose dimension depends only on k and d. Thus, we can use standard covering number bounds for norm balls in finite dimensions (compare, for instance, [63, Corollary 4.2.13] or [36, Example 1.10]) to get covering number bounds, depending only on k and d, for any single gate. Second, exploit the circuit structure together with subadditivity of the $1 \to 1$ norm (or even the diamond norm) on CPTP maps to construct from covering nets for the single gates a covering net for the whole PQC. For this second step, it is also useful to observe that, by Hölder's inequality for Schatten norms, $|\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) - \ell(\boldsymbol{\theta}'; \boldsymbol{x}, y)| \leq C_{\text{loss}} \left\| T^{\text{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}((|0\rangle\langle 0|)^{\otimes n}) - T^{\text{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}'}((|0\rangle\langle 0|)^{\otimes n}) \right\|_1$, so we can effectively ignore the loss observables in our covering number proofs when suitably taking the Lipschitz-type constant C_{loss} into account.

Given the relation between empirical covering numbers and the pseudo-dimension discussed in Theorem 3.3.2 and the pseudo-dimension bound of Theorem 4.2.1, the slightly superlinear scaling of empirical metric entropies with the number of trainable gates stated in Theorem 4.2.2 may not be surprising at first glance. Indeed, plugging the pseudo-dimension bound of $\lesssim \Gamma \log_2(\Gamma)$ into Theorem 3.3.2 leads to an empirical metric entropy bound of $\lesssim \Gamma \log_2(\Gamma) \log_2(1/\varepsilon)$. Indeed, while the bound in Theorem 4.2.2 provides the slightly better $\lesssim \Gamma \log_2(\Gamma/\varepsilon)$, the scaling in Γ, the relevant parameter of the PQC, is the same in both bounds. However, importantly, the complexity bounds of Theorem 4.2.2 apply to general PQCs, without restrictions on the encoding strategy or the measurements. Moreover, in [6, Theorem 9], we also show that the reasoning behind Theorem 4.2.2 extends to even more general PQCs, e.g., when multiple copies of a QML model are run in parallel, or when not only continuous parameters inside gates but also discrete structural parameters are optimized during training. While also the pseudo-dimension bounds of [1] can be generalized, e.g., to data re-uploading PQCs, these extensions are arguably more straightforward on the level of covering numbers. Nevertheless, the empirical covering number bounds of Theorem 4.2.2 do not constitute a strict generalization of the pseudo-dimension bound of Theorem 4.2.1 because, as discussed after Theorem 3.3.2, there is no general upper bound on the pseudo-dimension via empirical covering numbers.

Using the machinery of Section 3.2, the complexity measure upper bounds of Theorem 4.2.1 and 4.2.2 imply generalization guarantees. More concretely, starting from the pseudo-dimension bound, we can invoke [57, Corollary 3.3] or combine Theorems 3.2.9, 3.3.2, and 3.3.3. And given the empirical metric entropy bound, Theorems 3.2.9 and 3.3.3 together give generalization error bounds. These different pathways all lead to a generalization guarantee of the following flavor: A training data size scaling effectively as $\sim \Gamma \log_2(\Gamma)$ is, with high probability, sufficient for good generalization when using a variational QML model based on a PQC with Γ k-local trainable gates. While such guarantees become particularly useful for PQCs with especially few trainable gates, they already give an interesting insight for efficiently implementable variational QML models: If we use a PQC with $\text{poly}(n)$ k-local trainable gates for machine learning, $\text{poly}(n)$ training data points suffice to guarantee good generalization, with high probability.

A Quantum Process Tomography Perspective: In the discussion above, we have emphasized the perspective of variational QML. That is, we have interpreted PQCs in terms of the corresponding QML models, and we have used the complexity measure bounds from Theorems 4.2.1 and 4.2.2 to obtain generalization guarantees for them. At this point, we describe a somewhat different perspective on the above results, in the spirit of [1, Section 4.2]. Namely, if the above was a (quantum) machine learning perspective, we now take a more quantum information-theoretic perspective and interpret the results from the perspective of quantum process tomography.

It is well known that full quantum state and process tomography are information-theoretically expensive tasks, requiring a number of samples of the unknown state or process that grows exponentially in the number of qudits involved [110]. However, depending on the task at hand, full tomography might not be necessary. Theorems 4.2.1 and 4.2.2 and the corresponding generalization bounds tell us that a probably approximately correct variant of quantum process tomography for an unknown quantum circuit with Γ unknown k-local gates is possible already from roughly $\sim \Gamma \log_2(\Gamma)$ training samples. In particular, if the unknown circuit has $\mathrm{poly}(n)$ gates, then $\mathrm{poly}(n)$ samples are sufficient for PAC quantum process tomography. This can mean a significant improvement in sample size compared to the $\exp(n)$ samples required for full process tomography. Note: As both [1] and [6] also contain results similar to Theorems 4.2.1 and 4.2.2 that apply to variable-structure PQCs, we can obtain useful sample complexity bounds for PAC process tomography even without assuming the structure of the unknown quantum circuit to be known in advance.

The interpretation of the above as results about a relaxed version of quantum process tomography fits well into recent developments in quantum information research focused on learning classical representations of quantum objects from data. Notable papers in this direction include [111–123]. We can trace this line of research back to the PAC variant of quantum state tomography introduced by [124]. For comparison, while [124] proved that "pretty good quantum state tomography" of an unknown n-qubit state is possible from training data of size scaling linearly in n, Theorems 4.2.1 and 4.2.2 imply that "pretty good quantum process tomography" for an unknown quantum circuit with Γ k-local gates is possible from training data of size scaling slightly superlinearly in Γ.

Comparison to related work: We conclude this subsection by comparing the results of Theorems 4.2.1 and 4.2.2 and the corresponding generalization bounds with two selected related works. In [1], we used pseudo-dimension bounds as in Theorem 4.2.1 in combination with results of [57] to deduce generalization guarantees for learning PQCs in a realizable setting. Ref. [125] shows similar generalization bounds also in the agnostic setting. While not explicitly stated in [1], our pseudo-dimension bounds for parametrized quantum circuits imply empirical covering number bounds when combined with Theorem 3.3.2. In fact, the bound obtained through this combination recovers the scaling in the number of trainable gates in the PQC stated in [125, Theorem 5], extends this bound beyond unitary circuits to CPTP circuits, and, in contrast to the bound of [125, Theorem 5], is independent of the sample size.

A further work showing empirical covering number bounds for PQCs is [126]. In fact, the empirical metric entropy bounds proved in [126] are similar to those of Theorem 4.2.2. However,

the attention in [126] is restricted to unitary PQCs acting on quantum data, whereas our results in [6] apply also to PQCs for processing classical data and for PQCs consisting of general CPTP gates. Moreover, we significantly extend the reach of the proof technique to include more general PQC architectures, as already discussed above. Finally, our more careful application of methods from statistical learning theory in [6] leads to a quadratic improvement in the dependence of the generalization error bound on the number of trainable gates compared to [126, Theorem 2].
With this, we end our discussion of generalization guarantees for PQC-based variational QML models arising from properties of the trainable part of the PQC. We now turn our attention to how the choice of classical-to-quantum encoding strategy can influence generalization.

4.2.2 Generalization Guarantees Based on the Data-Encoding

Whereas the results from Subsection 4.2.1 apply for variational QML models independently of whether they are used on classical or quantum data, in this subsection, we focus on the case of classical data inputs. In this scenario, we have to decide how to encode classical data into the quantum circuit. Here, we investigate the impact of the choice of an encoding strategy on the generalization behavior of the QML model.

From PQCs to Generalized Trigonometric Polynomials: As pointed out by [127, 128], the choice of classical-to-quantum encoding is crucial to the expressive power of a PQC for processing classical data. With the perspective on PQCs put forward in [4, 127, 128], we can make this intuitively reasonable statement mathematically precise for a broad class of encoding strategies. Namely, for the input space $\mathcal{X} = [0, 2\pi)^d$ of d-dimensional vectors of angles, we consider data-reuploading PQCs as in Fig. 4.3 and additionally assume that all encoding gates are of the form $E_j : [0, 2\pi)^d \mapsto \mathcal{CPTP}_{d^{k_j}}$, $E_j(\boldsymbol{x})(\rho) = e^{-ix_{n_j}H_j}\rho e^{ix_{n_j}H_j}$ for some $k_j \in \{1, \ldots, n\}$ and $n_j \in \{1, \ldots, d\}$. That is, upon input $\boldsymbol{x}$, the j^{th} encoding gate implements the unitary channel on the associated k_j qudits obtained by evolving along the time-independent Hamiltonian H_j for time x_{n_j}. While more general encoding maps are mathematically possible, the above class already covers a variety of commonly used ansätze in which classical inputs are processed as rotation angles, among them the hardware-efficient ansatz [129]. For these encodings, we now study functions $f_{\boldsymbol{\theta}} \in \mathcal{F}^{\text{QMLM}}$, which by Eq. (4.1.1) can be written as $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \operatorname{tr}\left[M \cdot T^{\text{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}\left((|0\rangle\langle 0|)^{\otimes n}\right)\right]$ for some Hermitian n-qudit observable M.
By writing out the action of $T^{\text{QMLM}}_{\boldsymbol{x},\boldsymbol{\theta}}$ as dictated by the circuit structure and then iteratively expanding in the eigenbases of the respective encoding Hamiltonians, Refs. [4, 127, 128] show that any $f_{\boldsymbol{\theta}} \in \mathcal{F}^{\text{QMLM}}$ implemented by the PQC-based QML model can also be written as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}} \exp(-i\boldsymbol{\omega}\cdot\boldsymbol{x}), \tag{4.2.4}$$

for some coefficients $(c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega} \subseteq \mathbb{C}$ such that $\|f\|_\infty \leq \|M\|$. Here, $\boldsymbol{\omega}\cdot\boldsymbol{x}$ denotes the standard inner product between the vectors $\boldsymbol{\omega} \in \mathbb{R}^d$ and $\boldsymbol{x} \in [0, 2\pi)^d$, and $\Omega \subseteq \mathbb{R}^d$ is the set of admissible frequency vectors, which is determined by the spectra of the encoding Hamiltonians H_j. In contrast, the coefficients $(c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega}$ depend on the choice of parameters $\boldsymbol{\theta}$, the measurement M, and on the eigenbases of the encoding Hamiltonians H_j. We will use the terminology of [4]

and refer to Eq. (4.2.4) as the *generalized trigonometric polynomial (GTP)* representation of $f_{\boldsymbol{\theta}}$. Here, the "generalized" expresses that, when using Euler's formula to rewrite Eq. (4.2.4) in terms of trigonometric functions, their arguments are of the form $\boldsymbol{\omega} \cdot \boldsymbol{x}$ with $\boldsymbol{\omega}$ not necessarily a vector with integer entries.

In [4, Section 3], we give a concrete prescription for how to obtain the frequency set Ω from the encoding Hamiltonians H_j. While determining the exact coefficients $(c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega}$ when starting from the PQC structure is challenging, the above GTP representation allows us to upper bound the complexity of $\mathcal{F}^{\mathrm{QMLM}}$ by upper bounding the complexity of the superset of GTPs with frequency spectrum Ω and infinity norm bounded by $\|M\|$. That is, we can focus our attention on the class

$$\mathcal{F}^{\mathrm{GTP}} \coloneqq \left\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}} \exp(-i\boldsymbol{\omega}\cdot\boldsymbol{x}) \mid \|f\|_\infty \leq \|M\| \right\}. \tag{4.2.5}$$

Generalization Bounds for GTPs and PQCs: We give the following two complexity bounds for $\mathcal{F}^{\mathrm{GTP}}$ in terms of the size of the admissible frequency spectrum:

Theorem 4.2.3 (Rademacher Complexity and Metric Entropy Bounds for GTPs [4, Lemmas 3 and 9])**.** *Let $d \in \mathbb{N}_{\geq 1}$. Let $\mathcal{F}^{\mathrm{GTP}}$ be the hypothesis class defined in Eq. (4.2.5). Let $m \in \mathbb{N}_{\geq 1}$ and $\boldsymbol{x}_1, \ldots \boldsymbol{x}_m \in [0,2\pi)^d$.*

1. *The empirical Rademacher complexity of $\mathcal{F}^{\mathrm{GTP}}$ with respect to $S|_{\mathcal{X}} \coloneqq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ can be upper-bounded as*

$$\hat{\mathcal{R}}_{S|_{\mathcal{X}}}(\mathcal{F}^{\mathrm{GTP}}) \leq C \cdot (2\pi)^{\frac{d}{2}} \|M\| \cdot \sqrt{\frac{|\Omega| \log(|\Omega|)}{m}}, \tag{4.2.6}$$

   where $C > 0$ is some universal constant.

2. *Let $\varepsilon > 0$. The empirical metric entropy of $\mathcal{F}^{\mathrm{GTP}}$ with respect to $S|_{\mathcal{X}} \coloneqq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ can be upper-bounded as*

$$\log_2 \mathcal{N}(\mathcal{F}^{\mathrm{GTP}}, \|\cdot\|_{2,S|_{\mathcal{X}}}, \varepsilon) \leq C \cdot |\Omega| \log\left(\frac{(2\pi)^{\frac{d}{2}} \|M\| \cdot |\Omega|}{\varepsilon} \right), \tag{4.2.7}$$

   where $C > 0$ is some universal constant.

To prove the Rademacher complexity bound, we show how to interpret $\mathcal{F}^{\mathrm{GTP}}$ as a class of functions implemented by a simple classical neural network with a single hidden layer and with trigonometric activation functions. For such an architecture, we can then bound the empirical Rademacher complexity, here $|\Omega|$ acts as the number of neurons in the hidden layer. Our proof of the empirical metric entropy bound relies on the insight that $|\Omega|$ limits the effective dimensionality of $\mathcal{F}^{\mathrm{GTP}}$. We make this intuitive statement formal by showing how to lift covering nets from an object in a $|\Omega|$-dimensional normed real vector space to obtain covering nets for $\mathcal{F}^{\mathrm{GTP}}$.

Using the tools introduced in Chapter 3, namely Theorems 3.2.9 and 3.2.15, we can use the complexity bounds of Theorem 4.2.3 to derive generalization bounds for classes of GTPs. As the bounds in Theorem 4.2.3 crucially depend on $|\Omega|$, the size of the admissible frequency spectrum,

so do the resulting generalization error bounds. Namely, we obtain generalization guarantees of the following type: A training data size scaling effectively as $\sim |\Omega| \log(|\Omega|)$ is, with high probability, sufficient for good generalization when using a QML model with a GTP hypothesis class $\mathcal{F}^{\mathrm{GTP}}$ with frequency spectrum Ω. In this informal statement of the generalization bound, we focus on the dependence on $|\Omega|$. A more detailed statement, including the dependencies on $\|M\|$ and d, can be found in [4, Theorems 6 and 10].
While the generalization error bounds just discussed apply to classes of GTPs, together with our above observation the functions implemented by data re-uploading PQC-based QML models can be understood as GTPs, they imply generalization bounds for variational QML. Importantly, the $|\Omega|$-dependence on the level of GTPs becomes a dependence on the classical-to-quantum data-encoding strategy on the level of PQCs, since Ω is determined by the encoding Hamiltonians. Moreover, as we show in detail in [4, Section 6], for different natural choices of encoding Hamiltonians, $|\Omega|$ can be upper-bounded by an expression polynomial in the number of encoding Hamiltonians appearing in the PQC. For example, if we restrict the set of possible encoding Hamiltonians to Pauli strings, then $|\Omega| \leq \left(\frac{2N}{d} + 1\right)^d$, with N the number of encoding Hamiltonians. Thus, the generalization bounds proved in this subsection depend explicitly on architectural properties of the data-encoding strategy, such as which encoding Hamiltonians are used and how often they appear.

Outlook: Combining Different Generalization Guarantees: In this Chapter, we have described two different routes towards understanding the generalization behavior of PQC-based variational QML models, first via the trainable gates, then via the encoding gates. There is a natural strategy for combining the two bounds obtained in this way, namely via a simple union bound, as discussed in [4, Section 7]. In fact, given a countable number of generalization bounds for the same class, we can always combine them using a union bound. In the case of variational QML, such a reasoning effectively allows us to pick whichever generalization bound among the trainable-gate-based or the encoding-gate-based on is stronger. This implies that even a PQC-based model with many trainable gates can enjoy good generalization if it contains only few encoding gates and vice versa. In particular, as soon as at least one among the number of trainable gates or the number of encoding gates scales only polynomially with the number of qudits, then training data of size scaling polynomially in the number of qudits will suffice for good generalization.
As a final comment for this chapter, we want to emphasize that this union bound-based reasoning for obtaining generalization bounds that depend on both the trainable and the encoding part of a PQC is only possible a posteriori, once separate generalization bounds for trainable and encoding part are available. As such, this approach fails to take interactions between these two parts into account. This points to a path for tightening our generalization guarantees: Instead of studying separately the effects of the trainable gates and the encoding gates on the generalization performance of the variational QML model, we may want to take both of them into account at the same time. For example, studying the coefficients $(c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega}$ that can appear in the GTP representation of functions implemented by a PQC-based model in more detail could be one way of incorporating the trainable part into our reasoning for data re-uploading PQCs.

Chapter 5

Learning from Quantum Data

In Chapter 4, we presented variational quantum machine learning as one confluence of quantum information theory and statistical learning theory. In particular, we saw that variational quantum machine learning can be applied both to classical and quantum data. This chapter takes a broader perspective on problems of learning from quantum data and explores the training data requirements for two concrete tasks.

The chapter is subdivided into two parts. First, in Section 5.1, we consider a model of learning classical functions from quantum superposition examples and discuss the specific case of learning Boolean linear functions in more detail. Second, in Section 5.2, we turn our attention to tasks of learning quantum processes from quantum data, with a focus on a toy model for learning quantum state preparation procedures.

5.1 Learning Classical Concepts from Quantum Examples

A prominent line of research in quantum learning theory revolves around learning classical Boolean functions assuming quantum data access. While such data access can come in different forms, among them quantum membership queries [130–132] and quantum statistical queries [133], we focus on quantum superposition examples as introduced in [134] and refer the reader to the survey [135] for an overview over other approaches.

Classically, as discussed in Chapter 3, given a data-generating distribution $P \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$ for some input space $\mathcal{X}$ and output space $\mathcal{Y}$, a training example $(\boldsymbol{x}, y)$ in statistical learning theory is drawn at random according to the distribution P. If $\mathcal{X} = \{0,1\}^n$ for some $n \in \mathbb{N}_{\geq 1}$ and $\mathcal{Y} = \{0,1\}$, we take a *quantum superposition example* for P to be a pure quantum state of the form

$$\sum_{\boldsymbol{x} \in \{0,1\}^n} \sum_{y \in \{0,1\}} \sqrt{P(\boldsymbol{x}, y)}\, |\boldsymbol{x}, y\rangle \ , \tag{5.1.1}$$

a superposition of computational basis states [134]. A corresponding quantum training data set of size m is then given by the m^{th} tensor power of the state in Eq. (5.1.1). In particular, for a setting of realizable learning, where $P \in \mathrm{Prob}(\{0,1\}^n)$ is a probability distribution over n-bit

inputs and there is some (unknown) target function $f_* : \{0,1\}^n \to \{0,1\}$, the corresponding quantum superposition example is

$$|\psi_{P,f_*}\rangle \coloneqq \sum_{\boldsymbol{x}\in\{0,1\}^n} \sqrt{P(\boldsymbol{x})}\,|\boldsymbol{x}, f_*(\boldsymbol{x})\rangle \ , \tag{5.1.2}$$

and a quantum data set is again obtained by taking a tensor power of this state. Such quantum examples are at least as information-theoretically powerful as their classical counterparts, since we can generate a suitably randomly sampled classical example by performing a computational basis measurement on the state in Eq. (5.1.1) or (5.1.2). However, it is a priori unclear whether quantum superposition examples are strictly more powerful than classical examples and, if so, for which tasks. And even if the physical plausibility of such quantum data is still a matter of debate, compare the discussions in [136, 137], we can already investigate its potential and limitations mathematically.

In the case of distribution-independent PAC learning, both agnostic and realizable, the series of works [79, 138, 139] has shown that the sample complexities of learning from classical and from quantum examples differ by at most a constant factor. Both are characterized in essentially the same way by the VC dimension of the hypothesis class under consideration. Therefore, when trying to separate classical from quantum examples, we investigate distribution-dependent learning scenarios, in which the underlying distribution is known to the learner in advance. By now, there are several results about advantages, both information-theoretical and computational, of quantum over classical training data for distribution-dependent PAC learning [5, 11, 12, 14, 132, 134, 140–143]. Most of these works considered learning with respect to the uniform distribution, i.e., with $P(\boldsymbol{x}) = 1/2^n$ for all $\boldsymbol{x} \in \{0,1\}^n$. In the remainder of this section, we discuss how quantum Fourier sampling can serve as a useful tool for learning from quantum superposition examples and present the concrete task of learning Boolean linear functions in more detail. For both of these points, we will not restrict our attention to the uniform distribution, but allow for more general biased product distributions.

Quantum Fourier Sampling as a Subroutine for Quantum Learning: Since the pioneering works [144, 145], Fourier analysis has played an important role in the learning theory of Boolean functions. As nicely presented in [146, Chapter 3], recent decades have led to a variety of insights into the complexity of the Fourier spectrum for certain classes of Boolean functions, as well as into the resulting implications for learning theory. Here, we only partially review the basics of Fourier analysis for Boolean functions and sketch some of its applications for learning algorithms, with a particular emphasis on quantum learning.

Before beginning our short review, the material of which can be found in [146, Section 8.4], we recall that equivalently to the view of Boolean functions as mapping elements of $\{0,1\}^n$ to $\{0,1\}$, after the simple relabeling $0 \to 1$ and $1 \to -1$, we can regard them as mapping elements of $\{-1,1\}^n$ to $\{-1,1\}$. The latter is the perspective on Boolean functions taken in our presentation of Fourier analysis. We first define the underlying probability distributions that we use:

Definition 5.1.1 (Biased Product Distributions). *For a bias vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in [-1,1]^n$, we define the corresponding $\boldsymbol{\mu}$-biased product distribution on $\{-1,1\}^n$ via*

$$P_{\boldsymbol{\mu}}(\boldsymbol{x}) \coloneqq \prod_{i=1}^{n} \frac{1 + x_i \mu_i}{2}, \qquad \text{for } \boldsymbol{x} = (x_1, \ldots, x_n) \in \{-1,1\}^n. \tag{5.1.3}$$

Such a probability distribution now gives rise to an inner product on $\mathbb{R}^{\{-1,1\}^n}$, defined as

$$\langle f, g \rangle_{\boldsymbol{\mu}} \coloneqq \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{\mu}}}[f(\boldsymbol{x}) g(\boldsymbol{x})], \qquad \text{for } f, g : \{-1,1\} \to \mathbb{R}. \tag{5.1.4}$$

If we define, for $\boldsymbol{j} = (j_1, \ldots, j_n) \in \{0,1\}^n$, the function

$$\phi_{\boldsymbol{\mu},\boldsymbol{j}} : \{-1,1\}^n \to \mathbb{R}, \phi_{\boldsymbol{\mu},\boldsymbol{j}}(\boldsymbol{x}) = \prod_{i : j_i = 1} \frac{x_i - \mu_i}{\sqrt{1 - \mu_i^2}}, \tag{5.1.5}$$

then the set $\{\phi_{\boldsymbol{\mu},\boldsymbol{j}}\}_{\boldsymbol{j} \in \{0,1\}^n}$ forms an orthonormal basis for $(\mathbb{R}^{\{-1,1\}^n}, \langle \cdot, \cdot \rangle_{\boldsymbol{\mu}})$. We can now expand any Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ in terms of this ortonormal basis to obtain its $\boldsymbol{\mu}$-biased Fourier expansion. That is, we define the $\boldsymbol{\mu}$-biased Fourier coefficients of f as

$$\hat{f}_{\boldsymbol{\mu}}(\boldsymbol{j}) \coloneqq \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{\mu}}}[f(\boldsymbol{x}) \phi_{\boldsymbol{\mu},\boldsymbol{j}}(\boldsymbol{x})], \tag{5.1.6}$$

so that $f(\boldsymbol{x}) = \sum_{\boldsymbol{j} \in \{0,1\}^n} \hat{f}_{\boldsymbol{\mu}}(\boldsymbol{j}) \phi_{\boldsymbol{\mu},\boldsymbol{j}}(\boldsymbol{x})$ holds for all $\boldsymbol{x} \in \{-1,1\}^n$. We will refer to the collection of $\boldsymbol{\mu}$-biased Fourier coefficients as the $\boldsymbol{\mu}$-biased Fourier spectrum of f. If $\boldsymbol{\mu} = \mathbf{0} \in [-1,1]^n$ is the zero vector, which means that $P_{\boldsymbol{\mu}}$ is the uniform distribution over $\{-1,1\}^n$, we simplify the notation for Fourier coefficients as $\hat{f}(\boldsymbol{j}) \coloneqq \hat{f}_{\mathbf{0}}(\boldsymbol{j})$. Notice that, for $f : \{-1,1\}^n \to \{-1,1\}$, Parseval's identity becomes

$$\sum_{\boldsymbol{j} \in \{0,1\}^n} \left(\hat{f}_{\boldsymbol{\mu}}(\boldsymbol{j}) \right)^2 = \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{\mu}}}[f(\boldsymbol{x})^2] = 1, \tag{5.1.7}$$

So, the squares of the $\boldsymbol{\mu}$-biased Fourier coefficients form a probability distribution over the Boolean hypercube $\{0,1\}^n$.
One way of connecting the notions of Fourier analysis introduced above to learning-theoretic questions goes through the following elementary observation:

Lemma 5.1.2 ([144, Lemma 9]). *If $f : \{-1,1\}^n \to \{-1,1\}$ and $g : \{-1,1\} \to \mathbb{R}$, then*

$$\mathbb{P}_{\boldsymbol{x} \sim P_{\boldsymbol{\mu}}}[f(\boldsymbol{x}) \neq \operatorname{sgn}(g(\boldsymbol{x}))] \leq \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{\mu}}}[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2] = \sum_{\boldsymbol{j} \in \{0,1\}^n} \left(\hat{f}_{\boldsymbol{\mu}}(\boldsymbol{j}) - \hat{g}_{\boldsymbol{\mu}}(\boldsymbol{j}) \right)^2. \tag{5.1.8}$$

Here, sgn *denotes the sign function, where* 0 *is assigned sign* $+1$.

As a consequence of this Lemma, whenever we are able to approximate the $\boldsymbol{\mu}$-biased Fourier coefficients of an unknown Boolean function f, then we obtain a corresponding PAC approximation to f with respect to the 0-1-loss and the underlying distribution $P_{\boldsymbol{\mu}}$. In general, without prior assumptions on the Fourier spectrum of f, this approach towards learning f requires us to approximate an exponentially-in-n large number of Fourier coefficients, and thus does not lead to

an information-theoretically or computationally efficient learning strategy. If, however, we know a priori that the set of non-negligible Fourier coefficients has small cardinality, the approach can be more fruitful, especially if that set is known in advance.
Quantum Fourier sampling can serve as a subroutine for identifying the relevant Fourier coefficients of an unknown Boolean function, assuming quantum example access. At the basis of this approach lies the biased quantum Fourier transform, extending the standard quantum Fourier transform [13]:

Definition 5.1.3 (Biased Quantum Fourier Transform [14]). *For $n \in \mathbb{N}_{\geq 1}$ and a bias vector $\boldsymbol{\mu} \in (-1,1)^n$, the n-qubit $\boldsymbol{\mu}$-biased quantum Fourier transform $H^n_{\boldsymbol{\mu}}$ acts on a computational basis state $|\boldsymbol{x}\rangle$ with $\boldsymbol{x} \in \{-1,1\}^n$ as*

$$H^n_{\boldsymbol{\mu}} |\boldsymbol{x}\rangle \coloneqq \sum_{\boldsymbol{j}\in\{0,1\}^n} \sqrt{P_{\boldsymbol{\mu}}(\boldsymbol{x})}\phi_{\boldsymbol{\mu},\boldsymbol{j}}(\boldsymbol{x})|\boldsymbol{j}\rangle\,. \tag{5.1.9}$$

Armed with the unitary $H^n_{\boldsymbol{\mu}}$, we can describe the procedure for $\boldsymbol{\mu}$-biased quantum Fourier sampling: Given a quantum example $|\psi_{P_{\boldsymbol{\mu}},f}\rangle \coloneqq \sum_{\boldsymbol{x}\in\{-1,1\}^n} \sqrt{P_{\boldsymbol{\mu}}(\boldsymbol{x})}\,|\boldsymbol{x}, f(\boldsymbol{x})\rangle$ of a function $f : \{-1,1\}^n \to \{0,l\}$ – this is the same as in Eq. (5.1.2) up to a relabeling of inputs – we apply $H^n_{\boldsymbol{\mu}}$ to the first n-qubits and $H \coloneqq H^1_0$ to the last qubit, and then measure all $n+1$ qubits in the computational basis. When following this prescription, the output is as follows:

Lemma 5.1.4 ([14, Lemma 3]). *When performing $\boldsymbol{\mu}$-biased quantum Fourier sampling on a quantum example $|\psi_{P_{\boldsymbol{\mu}},f}\rangle \coloneqq \sum_{\boldsymbol{x}\in\{-1,1\}^n} \sqrt{P_{\boldsymbol{\mu}}(\boldsymbol{x})}\,|\boldsymbol{x}, f(\boldsymbol{x})\rangle$ of a function $f : \{-1,1\}^n \to \{0,l\}$, we observe the measurement outcome 1 for the last qubit with probability $\frac{1}{2}$. Conditioned on that event, we observe outcome $\boldsymbol{j} \in \{0,1\}^n$ for the first n qubits with probability $(\hat{g}_{\boldsymbol{\mu}}(\boldsymbol{j}))^2$, where $g : \{-1,1\}^n \to \{-1,1\}$ is defined as $g(\boldsymbol{x}) = (-1)^{f(\boldsymbol{x})}$.*

Hence, as its name suggests, $\boldsymbol{\mu}$-biased quantum Fourier sampling indeed allows us to sample from the probability distribution formed by the squares of $\boldsymbol{\mu}$-biased Fourier coefficients of the unknown function, assuming access to quantum superposition examples. In particular, repeating this sampling with multiple quantum examples, we can identify the non-negligible Fourier coefficients of the unknown function, thereby achieving an important first step for a Fourier-based learning strategy. Notably, classical Fourier sampling from an unknown Boolean function via the Goldreich-Levin algorithm [147] works on the basis of membership access, whereas quantum example access is sufficient for quantum Fourier sampling.

Learning Boolean Linear Functions From Quantum Examples: Learning strategies based on quantum Fourier sampling have proven useful for a variety of function classes, among them disjunctive normal forms [14, 134, 140] and Fourier-sparse functions [132]. In addition, and interestingly from a cryptographic perspective, Fourier-based quantum learning has successfully been used for exactly learning Boolean linear functions with respect to the uniform distribution [13], even from noisy quantum data [11, 12, 142]. Here, we take exact learning of Boolean linear functions as an example to analyze the possible advantage of quantum over classical examples as the underlying distribution deviates from uniformity.

A Boolean linear function described by a bit string $\boldsymbol{a} \in \{0,1\}^n$ is defined as

$$f^{(\boldsymbol{a})} : \{-1,1\}^n \to \{0,1\},\ f^{(\boldsymbol{a})}(\boldsymbol{x}) \coloneqq \sum_{i=1}^{n} a_i \frac{1-x_i}{2} \pmod 2. \tag{5.1.10}$$

When the underlying distribution is uniform, the Fourier spectrum of a Boolean linear function is particularly simple. Namely, for $g^{(\boldsymbol{a})} : \{-1,1\}^n \to \{-1,1\}$ defined as $g^{(\boldsymbol{a})}(\boldsymbol{x}) = (-1)^{f^{(\boldsymbol{a})}(\boldsymbol{x})}$ we have $\hat{g}^{(\boldsymbol{a})}(\boldsymbol{j}) = \delta_{\boldsymbol{a},\boldsymbol{j}}$ for $\boldsymbol{j} \in \{0,1\}^n$. In particular, as observed in [13], a single successful run of unbiased quantum Fourier sampling on a quantum example state of an unknown Boolean linear function suffices to identify the unknown string $\boldsymbol{a}$ exactly. Here, we speak of a successful run if the measurement outcome observed for the last qubit is 1. As a single run is successful with probability $1/2$, this observation implies:

Lemma 5.1.5. *Let $\delta \in (0,1)$. In the case of no bias, that is $\boldsymbol{\mu} = \mathbf{0} \in [-1,1]^n$, if $\boldsymbol{a} \in \{0,1\}^n$ is an unknown n-bit string, $\mathcal{O}(\log_2(1/\delta))$ quantum examples $|\psi_{P_{\mathbf{0},f^{(\boldsymbol{a})}}}\rangle$ of $f^{(\boldsymbol{a})}$ suffice to exactly identify $\boldsymbol{a}$, with probability $\geq 1-\delta$.*

Interestingly, the upper bound on the sufficient training data size in Lemma 5.1.5 is independent of n. In contrast, solving the analogous classical learning problem – that is, exactly identifying an unknown $\boldsymbol{a} \in \{0,1\}^n$ from classical training examples $(\boldsymbol{x}_1, f^{(\boldsymbol{a})}(\boldsymbol{x}_1)), \ldots, (\boldsymbol{x}_m, f^{(\boldsymbol{a})}(\boldsymbol{x}_m))$, with the $\boldsymbol{x}_i$ drawn i.i.d. from the uniform distribution over $\{-1,1\}^n$ – requires a training data size of $m = \Omega(n)$. This well-known fact can, e.g., be proven through information-theoretic considerations, compare [5, Theorem 5]. Therefore, for the problem of exactly learning an unknown Boolean linear function with respect to the uniform distribution, quantum superposition examples provide an information-theoretic advantage over classical examples: Whereas the classical sample complexity depends linearly on n, the quantum sample complexity is independent of n. This advantage quite directly translates into a similar one in terms of computational complexities. And while the classical version of the problem can be solved using $\mathcal{O}(n)$ examples and thus efficiently, this simple quantum speedup and its relatives can offer us further insight into the power of quantum examples.

A variant of the problem of learning linear functions, the so-called "Learning With Errors (LWE)" problem, or more precisely its assumed hardness, is important for cryptographic purposes [148]. Classically, in LWE for bits – in which case the problem is also referred to as "Learning Parity With Noise (LPN)" –, training data takes the form $\{(\boldsymbol{x}_i, f^{(\boldsymbol{a})}(\boldsymbol{x}_i) + e_i)\}_{i=1}^m$, where the $\boldsymbol{x}_i$ are drawn i.i.d. uniformly at random from $\{-1,1\}^n$, $\boldsymbol{a} \in \{0,1\}^n$ is an unknown (possibly also randomly chosen) bit string to be identified, and the e_i are noise terms drawn i.i.d. from some error distribution over $\{0,1\}$. While classical algorithms for LPN and LWE with sub-exponential sample and/or time complexity have been proposed [149–151], no polynomial-time algorithm is known. In contrast, as demonstrated by [11, 12, 142], even noisy quantum algorithms still allow for simple efficient Fourier-based quantum learning algorithms. Here, in the interest of brevity of presentation, we reproduce an informal statement from [12] in a simplified form:

Theorem 5.1.6 ([12, Main Result (Informal)])**.** *For error distributions used in cryptographic schemes, and for any $\delta \in (0,1)$, there exists a Fourier-based quantum learning algorithm that*

solves LWE with probability $\geq 1-\delta$ *using* $\mathcal{O}\left(n \log_2\left(1/\delta\right)\right)$ *noisy quantum examples and that runs in time* $\mathrm{poly}\left(n, \log_2\left(1/\delta\right)\right)$.

Thus, when adding noise the quantum examples, quantum Fourier sampling still proves to be a powerful tool for learning Boolean linear functions, even if the sample and time complexity bounds now become n-dependent.
In [5], we explored how the sample and time complexities, both for the noiseless and for the noisy case, change if the underlying distribution is no longer uniform but can be a general biased product distribution. Importantly, if the underlying distribution is biased, the correspondingly biased Fourier spectrum of a linear function is no longer a simple delta function. However, as we show in [5, Lemma 4], also the biased Fourier spectrum has a distinct structure, which can be exploited to generalize the Bernstein-Vazirani algorithm to the biased case, compare [5, Algorithm 2]. If the bias is small enough, this leads to a quantum learning algorithm that essentially recovers the n-independent sample complexity upper bound of $\mathcal{O}\left(\log_2\left(1/\delta\right)\right)$ of the unbiased case [5, Theorem 3]. Even for larger bias, as long as no input bit is fully biased to ± 1, the generalized version of Bernstein-Vazirani can serve as the basis of a quantum exact learner for linear functions, then again with a sample complexity upper bound polylogarithmic in n [5, Theorem 4]. In the case of large bias, such a polylogarithmic dependence cannot be avoided [5, Theorem 6]. For small bias, we additionally show that linear functions can be learned from noisy quantum examples with a sample complexity that depends polylogarithmically on n, under a certain structural assumption on the noise model [5, Appendix A.1]. This strengthens the case for noise-robustness of Fourier-based quantum learning, providing a better bound than Theorem 5.1.6 for more general distributions under slightly more restrictive noise models. Moreover, our results in [5] provide quantitative insight into how the power of quantum superposition examples depends on the superposition weights.

5.2 Learning State Preparation Procedures from Quantum Data

In the framework of Section 5.1, the training data was quantum, but the object to be learned was still a classical function. This section now considers learning settings in which the object to be learned is itself quantum, which makes it natural to let the data and learning algorithms be quantum as well.
To describe such a quantum learning problem, we imagine an unknown physical process as the object to be learned. According to the mathematical framework introduced in Chapter 2, we can thus think of an unknown CPTP map as the target object of learning. We can now consider different scenarios of obtaining information about the unknown process. For example, in the spirit of Chapter 3, we might describe models of learning a CPTP map from data consisting of randomly sampled quantum input-output examples, as in [3, 152]. However, we can also consider scenarios of learning from queries, in which a learner can query the unknown map multiple times and we allow queries and the processing thereof to be adaptively chosen depending on previous steps of the protocol. Scenarios like this were explored in [153–158].
These different formulations of quantum learning models raise new questions beyond those common in classical learning theory. On the one hand, we have to physically justify the form of

training data. For instance, as discussed in [152], due to the no-cloning theorem for quantum states, not in all scenarios of randomly sampled examples is it reasonable to assume that a learner has access to both quantum input and quantum output states at the same time. To avoid this issue, [3, 152] take examples consisting of a classical description of an input state and an actual quantum copy of the corresponding output state. On the other hand, novel quantum learning scenarios also allow for qualitatively different kinds of learning algorithms. Naturally, simple semi-classical procedures, in which each quantum example is measured separately and the observed measurement outcomes are processed classically, are possible. However, when learning from randomly sampled quantum examples, we can also allow for quantum learners that measure multiple quantum examples simultaneously. And if we have query access to an unknown CPTP map, quantum information theory also allows for quantum learning procedures that can potentially make use of entanglement through coherent processing of quantum data, as emphasized in [153, 154]. Both of these kinds of quantum processing hinge on the assumption that the learner has access to a quantum memory, the importance of which was pointed out in [155–157]. Given this, for inherently quantum learning problems, it is of particular interest whether the qualitative differences between learning strategies, such as coherent versus incoherent access or quantum versus classical memory, translate to quantitative differences in terms of sample and/or computational complexity. Effectively, this is the question of whether quantum learning algorithms have an advantage over classical learners for naturally quantum learning tasks. Before discussing this question by way of a concrete example, we shortly review some recent results in this line of research.

Independently, both [153] and [154] presented a general framework for investigating the power of coherent quantum data access learning from physical experiments. [153, Theorem 1] presents a limitation on the potential improvement in query complexity when going from incoherent to coherent access for tasks in which the focus is on a kind of average-performance. However, [153, Theorems 2 and 3] show an exponential sample complexity advantage of coherent over incoherent processing for predicting expectation values of Pauli observables in a worst-case model. And [154, Theorem 1] establishes a similar exponential advantage for a specific problem of distinguishing two different physical processes. This line of work was continued by [155], who highlighted the relevance of the availability of quantum memory. In particular, [155] proposed a novel proof strategy leading to generalizations of the separation results of [153, 154] and to similarly strong separations for further tasks. In addition, [156] analyzed separations arising from differences in the number of quantum copies that can be measured simultaneously. Finally, the results of [157] demonstrate that machine learning models enhanced with coherent access to data from quantum experiments and with a quantum memory can significantly outperform their (semi-)classical counterparts already in the NISQ era.

Learning State Preparation Procedures – A Toy Model: To complement the general deliberations above, here we investigate a concrete task of learning from quantum data, with the goal of determining its sample complexity and an optimal learning algorithm. As a first step in "quantizing" a classical learning problem, where both the inputs and outputs are classical, we consider learning maps with classical input and quantum output. Physically, such a map

corresponds to a state preparation procedure: Given a choice of classical parameters, which we can for example think of as modifiable elements in an experimental setup, a corresponding quantum state is prepared. The goal of a quantum learner in such a model is then, given training data consisting of classical inputs and the corresponding quantum output states, to find a state preparation procedure that mimics the action of the unknown one. Next we describe a toy model for such a learning task in more detail and characterize its sample complexity.

Recall that in classical binary classification, the output space is $\mathcal{Y} = \{0, 1\}$ and a possible loss function is the 0-1-loss. To define our quantum version of binary classification, we instead consider an output space $\mathcal{Y} = \{\sigma_0, \sigma_1\}$, where $\sigma_0, \sigma_1 \in \mathcal{S}(\mathbb{C}^d)$ are qudit quantum states. The input space is some classical space $\mathcal{X}$. As the target space is still a binary one, the 0-1-loss is again a reasonable choice of loss function. For our purposes, it is natural to rescale the loss according to the distinguishability of the two label states as measured by their trace distance. That is, we take as loss function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0},\ \ell(\rho_1, \rho_2) \coloneqq \frac{\|\sigma_0 - \sigma_1\|_1}{2} \cdot \delta_{\rho_1,\rho_2} = \frac{\|\rho_1 - \rho_2\|}{2}\,. \tag{5.2.1}$$

Therefore, for a data distribution $P \in \mathrm{Prob}(\mathcal{X} \times \{\sigma_0, \sigma_1\})$ and a hypothesis $h : \mathcal{X} \to \{\sigma_0, \sigma_1\}$, the expected risk is

$$R(h) = \int_{\mathcal{X}\times\mathcal{Y}} \frac{\|\rho - h(\boldsymbol{x})\|_1}{2}\, \mathrm{d}P(\boldsymbol{x}, \rho)\,. \tag{5.2.2}$$

And the empirical risk of h on a training data set $S = \{(\boldsymbol{x}_i, \rho_i)\}_{i=1}^m \subseteq \mathcal{X} \times \{\sigma_0, \sigma_1\}$ is given as

$$\hat{R}_S(h) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \frac{\|\rho_i - h(\boldsymbol{x}_i)\|_1}{2}\,. \tag{5.2.3}$$

The above setup strongly resembles that of classical statistical learning established in Chapter 3. The crucial difference lies in the fact that the quantum label states ρ_i in a training data set $S = \{(\boldsymbol{x}_i, \rho_i)\}_{i=1}^m$ are provided as actual quantum states. In particular, given such data, a learner obtains a single quantum copy of each ρ_i. If σ_0 and σ_1 are not perfectly distinguishable, such a single copy is not sufficient to discern with certainty whether ρ_i equals σ_0 or σ_1. Instead, a learner has to process the quantum training data and then extract information through a measurement.

As the main result of [3], we characterize the sample complexity of this task of binary classification with classical instance and quantum labels. For agnostic learning, we obtain the following:

Theorem 5.2.1 ([3, Corollary 1]). *Let $\sigma_0, \sigma_1 \in \mathcal{S}(\mathbb{C}^d)$ be distinct. Let $\mathcal{F} \subseteq \{\sigma_0, \sigma_1\}^{\mathcal{X}}$ be a non-trivial hypothesis class. Define $\tilde{\mathcal{F}} \coloneqq \{\tilde{f} : \mathcal{X} \to \{0, 1\} \mid \exists f \in \mathcal{F} :\ f(x) = \sigma_{\tilde{f}(x)}\ \forall x \in \mathcal{X}\}$, and assume that $\mathrm{VCdim}(\tilde{\mathcal{F}}) < \infty$. Let $\ell : \{\sigma_0, \sigma_1\} \times \{\sigma_0, \sigma_1\} \to \mathbb{R}_{\geq 0}$ be the trace distance loss defined as $\ell(\rho_1, \rho_2) \coloneqq \frac{\|\rho_1 - \rho_2\|}{2}$. Then, for every $P \in \mathrm{Prob}(\mathcal{X} \times \{\sigma_0, \sigma_1\})$ and for every $\varepsilon \in (0, \|\sigma_0 - \sigma_1\|/8)$ and $\delta \in \left(0, 1 - \frac{1}{4}\log_2(\frac{1}{4}) - \frac{3}{4}\log_2(\frac{3}{4})\right)$, a sample size*

$$m = m(\varepsilon, \delta) = C \cdot \frac{\mathrm{VCdim}(\tilde{\mathcal{F}}) + \ln(1/\delta)}{\varepsilon^2}\,, \tag{5.2.4}$$

where $C > 0$ is a universal constant, is sufficient to guarantee: With probability $\geq 1-\delta$ over the choice of a training data set $S = \{(\boldsymbol{x}_i, \rho_i)\}_{i=1}^m$ consisting of m examples drawn i.i.d. according to P,

$$\sup_{h\in\mathcal{F}} |R(h) - \hat{R}_S(h)| \leq \varepsilon\,. \tag{5.2.5}$$

In addition, a sample size as in Eq. (5.2.4) (but with a different constant prefactor) is also necessary.

This agrees with the sample complexity for agnostic classical binary classification, compare Eq. (3.2.3). In the realizable case, we show in [3] that a sample size of

$$m = m(\varepsilon,\delta) = C\cdot \frac{\mathrm{VCdim}(\tilde{\mathcal{F}}) + \ln(1/\delta)}{\varepsilon(1-2\max\{\mathrm{tr}[E_0\sigma_1], \mathrm{tr}[E_1\sigma_0]\})^2}\,, \tag{5.2.6}$$

with $\{E_0, E_1\}$ the Holevo-Helstrom measurement achieving the optimal success probability in distinguishing σ_0 from σ_1, suffices for (improper) PAC learning, and that a sample size of

$$m = m(\varepsilon,\delta) = C\cdot \frac{\mathrm{VCdim}(\tilde{\mathcal{F}}) + \ln(1/\delta)}{\varepsilon} \tag{5.2.7}$$

is also necessary. Thus, despite the quantum nature of the labels, the sample complexities for agnostic and realizable PAC binary classification remain essentially unchanged compared to the classical case, they are again determined by the VC dimension of the hypothesis class.
To prove the sample complexity upper bounds, in [3] we describe semi-classical learning procedures for which the sample sizes are sufficient, proceeding as follows: First, the learner measures each quantum label separately using the Holevo-Helstrom measurement for $\{\sigma_0, \sigma_1\}$, thereby producing a classical training data set $\tilde{S}$ with labels in $\{0,1\}$. We can interpret $\tilde{S}$ as a noisy version of a "true" training data set, with the noise induced by the measurement. Second, the learner uses $\tilde{S}$ as input to a classical algorithm for learning from noisy data. In the agnostic case, the latter is simply empirical risk minimization, albeit with a modified loss function to account for the noise. For the realizable scenario, we combine techniques of [51, 159] to develop an information-theoretically optimal classical learner from noisy data. The sample complexity lower bounds are established by relating the respective learning problem to a quantum state discrimination task, the sample complexity of which we analyze information-theoretically.
The proofs of these results establish an additional insight beyond the bounds themselves. Namely, they show that no coherent processing of the quantum data is needed to achieve the optimal sample complexity for binary classification with classical instances and quantum data. For this specific task, this strengthens a result of [153], which for our purposes guarantees that the optimal achievable sample complexities of coherent and incoherent procedures for binary classification with classical instances and quantum labels differ at most by a factor polynomial in $\log_2(d)$. Moreover, our proofs demonstrate that it is not necessary for a quantum learner to receive a whole classical-quantum training data set at once, the examples can also be provided one after the other if the learner stores each measurement result in a classical memory. In the language of [155], no quantum memory is needed to achieve information-theoretic optimality here. We

note, however, that these results depend on our implicit assumption that a quantum learner knows a priori which label states σ_0 and σ_1 can appear.

In this chapter, we have demonstrated that, beyond the quantum computing-based models for classical machine learning discussed in Chapter 4, quantum information theory allows for inherently quantum learning problems. Interestingly, recent works such as [155, 157] pose tasks of learning from data generated through quantum experiments as promising candidates for information-theoretic and computational advantages of quantum over classical computing. However, as a consequence of the results of [79] and [153], an information-theoretic quantum advantage is not automatically possible, but the potential for it depends on different features of the learning problem, such as distribution-independent versus distribution-dependent or average-case versus worst-case. The results of this chapter add to this in two ways: On the one hand, in learning from quantum superposition examples, the underlying distribution influences the possible quantum advantage quantitatively. On the other hand, in learning simple state preparation procedures, the possible advantage of coherent quantum over incoherent semi-classical learning strategies can vanish.

Chapter 6

Quantum (Non-)Markovianity

In Chapters 4 and 5, we have discussed different questions of learning from data in the framework of quantum information theory. For this learning-theoretic perspective, we implicitly presuppose the usefulness of available data for making the predictions of interest. However, when the goal is to predict the future, data about the present-to-future evolution does not necessarily suffice.

As an example, consider the following thought experiment: At time $t = -2$, a bipartite product input quantum state $\rho_{AB} = \rho_A \otimes \tilde{\rho}_B$ is provided. Next, at $t = -1$, a measurement $\{E_i\}_i$ is performed on the B-subsystem, the post-measurement state is discarded, and the observed outcome i_* is recorded in a classical memory that the learner cannot access. Thus, the quantum state at the time $t = 0$, viewed from the perspective of the learner, is ρ_A. Finally, at $t = 1$, depending on the outcome i_* stored in the classical memory, a quantum channel T_{i_*} is applied to the A-system, leading to an output state $T_{i_*}(\rho_A)$. A learner, classical or quantum, that has access only to snapshots of the A-subsystem at times $t = 0$ and $t = 1$, sees data consisting of pairs of the form $(\rho_A, T_{i_*}(\rho_A))$. Now, if in a new run of the experiment the learner is confronted with an "input" state σ_A at the present time $t = 0$, she cannot reliably predict the corresponding future "output" state at $t = 1$, because the correct output depends on information lying further in the past than accessible through the data.

In this example, data about the present-to-future evolution is not enough to make predictions about the future. The use of a classical memory in the overall evolution from $t = -2$ to $t = 1$ leads to a direct dependence of the future ($t = 1$) on the past ($t = -2, -1$), thereby limiting the usefulness of information about the present-to-future ($t = 0$ to $t = 1$) evolution. This problem does not occur if we consider a memoryless evolution, in which the future depends on the past only through the present. In this sense, memorylessness, also called Markovianity, is important for guaranteeing that certain tasks of extracting information from data in quantum experiments are well-posed.

Motivated by this perspective on Markovianity as potential underlying justification for the usefulness of data in quantum experiments, in Section 6.1, we review finite-dimensional quantum dynamical semigroups as a strong notion of Markovianity for the continuous-time evolution of a quantum system. Next, we discuss a relaxation of this notion in terms of infinitesimal Markovian divisibility in Section 6.2. We conclude the chapter by extending the framework of quantum dynamical semigroups to quantum superchannels in Section 6.3.

6.1 Quantum Dynamical Semigroups

Continuous one-parameter semigroups serve as a traditional mathematical framework for time-homogeneous Markovian evolutions in continuous time:

Definition 6.1.1 (Continuous One-Parameter Semigroup)**.** *A family* $(T_t)_{t\geq 0}$ *of linear maps* $T_t \in \mathcal{B}(\mathbb{C}^d)$ *forms a* continuous one-parameter semigroup *if*

(i) $T_{t+s} = T_t T_s$ *for all* $s, t \geq 0$,

(ii) $T_0 = \mathbb{1}_d$, *and*

(iii) the map $\mathbb{R}_{\geq 0} \ni t \mapsto T_t \in \mathcal{B}(\mathbb{C}^d)$ *is continuous.*

In Definition 6.1.1, property (i) captures both Markovianity and time-homogeneity. It describes a Markovian evolution because, for any $0 \leq s \leq t$ and any input $x \in \mathbb{C}^d$, we can obtain $T_t(x)$, the output after time t, by processing $T_s(x)$, the output after time s, via the map T_{t-s}. And it describes a time-homogeneous evolution in the sense that the evolution from time s to time $t \geq s$ depends only on the time difference $t - s$. Properties (ii) and (iii) formalize the physically motivated intuitions that no non-trivial evolution takes place at time 0 and that the evolution depends continuously on time

Note also that, in property (iii) of Definition 6.1.1, we define continuity with respect to the topology on $\mathcal{B}(\mathbb{C}^d)$ induced by the operator norm, the topology of uniform convergence, (and with respect to the standard topology on $\mathbb{R}$). From a functional analytic perspective, strong convergence gives rise to a natural alternative topology on $\mathcal{B}(\mathbb{C}^d)$. In finite dimensions, the notions uniform and strong convergence coincide. Thus, we can work with uniform convergence without much need for justification. We will comment on generalizations to infinite dimensions only shortly towards the end of this section.

A crucial tool in studying continuous one-parameter semigroups are their *generators*:

Theorem 6.1.2 (Generators of Continuous One-Parameter Semigroups)**.** *Let* $(T_t)_{t\geq 0}$ *be a continuous one-parameter semigroup of linear maps* $T_t \in \mathcal{B}(\mathbb{C}^d)$. *Then, the map* $\mathbb{R}_{>0} \ni t \mapsto T_t \in \mathcal{B}(\mathbb{C}^d)$ *is differentiable. Moreover, there exists a linear map* $L \in \mathcal{B}(\mathbb{C}^d)$ *such that* $T_t = e^{tL}$ *for all* $t \geq 0$. *We call* L *the* generator *of* $(T_t)_{t\geq 0}$.

Proof sketch. Continuity of $t \mapsto T_t$, together with $T_0 = \mathbb{1}_d$ being invertible, by openness of the set of invertible linear maps in $\mathcal{B}(\mathbb{C}^d)$, equipped with the operator norm, implies that $M_\tau := \int_0^\tau T_s \, \mathrm{d}s$ is invertible for $\tau > 0$ small enough. And since M_τ is defined in terms of an integral, the map $\tau \mapsto M_\tau$ is differentiable. Now, observing that, by property (i), T_t can be rewritten as $T_t = M_\tau^{-1}(M_{t+\tau} - M_t)$, $t \mapsto T_t$ is differentiable as a composition of differentiable maps.

Moreover, this representation of T_t, via the fundamental theorem of calculus and property (i) in Definition 6.2.2, implies that $\frac{\mathrm{d}}{\mathrm{d}t} T_t = M_\tau^{-1}(T_\tau - \mathbb{1}_d)T_t$. Hence, defining $L := M_\tau^{-1}(T_\tau - \mathbb{1}_d)$, the continuous one-parameter semigroup satisfies the differential equation $\frac{\mathrm{d}}{\mathrm{d}t} T_t = L T_t$ with initial condition $T_0 = \mathbb{1}_d$. This has the unique solution $T_t = e^{tL}$, $t \geq 0$. □

Theorem 6.1.2 and its proof show us an alternative perspective on continuous one-parameter semigroups $(T_t)_{t\geq 0}$ in terms of a differential equation $\frac{\mathrm{d}}{\mathrm{d}t} T_t = L T_t$, $T_0 = \mathbb{1}_d$, where L generates

the semigroup. On this level, the Markovianity of the evolution is reflected by the fact that $\frac{\mathrm{d}}{\mathrm{d}t}T_t$ depends only on T_t, not on earlier times. Similarly, the time-independence of the generator L captures the time-homogeneity of the evolution.
So far, we have considered continuous one-parameter semigroups for general linear maps. Now, we focus on *quantum dynamical semigroups*, the case particularly relevant when considering the evolution of quantum systems:

Definition 6.1.3 (Quantum Dynamical Semigroups)**.** *A one-parameter family $(T_t)_{t\geq 0}$ of linear maps $T_t \in \mathcal{B}\left(\mathcal{B}(\mathbb{C}^d)\right)$ forms a* quantum dynamical semigroup *in the Schrödinger picture if*

(i) T_t *is CPTP for all* $t \geq 0$,

(ii) $T_{t+s} = T_t T_s$ *for all* $s, t \geq 0$,

(iii) $T_0 = \mathrm{id}_{\mathcal{B}(\mathbb{C}^d)}$, *and*

(iv) the map $\mathbb{R}_{\geq 0} \ni t \mapsto T_t \in \mathcal{B}\left(\mathcal{B}(\mathbb{C}^d)\right)$ *is continuous.*

In other words, in the Schrödinger picture, a quantum dynamical semigroup is a continuous one-parameter semigroup $(T_t)_{t\geq 0}$ of linear CPTP maps $T_t \in \mathcal{B}\left(\mathcal{B}(\mathbb{C}^d)\right)$. Naturally, to obtain the Heisenberg picture analogon of this definition, we have to replace CPTP by CPU.
By Theorem 6.1.2, we know that we can understand quantum dynamical semigroups in terms of their generators. In a seminal result for the study of Markovian quantum evolutions, these generators have been fully characterized:

Theorem 6.1.4 (GKLS/Lindblad Generators [15, 16])**.** *A linear map $L \in \mathcal{B}\left(\mathcal{B}(\mathbb{C}^d)\right)$ is the generator of a continuous one-parameter semigroup of CP maps if and only if it can be written as*

$$L(\rho) = \Phi(\rho) - K\rho - \rho K^\dagger\,, \tag{6.1.1}$$

where $\Phi \in \mathcal{B}(\mathcal{B}(\mathbb{C}^d))$ is a CP map and $K \in \mathcal{B}(\mathbb{C}^d)$ is arbitrary. Moreover, L is the generator of a quantum dynamical semigroup (in the Schrödinger picture) if and only if it can be written in the form of Eq. (6.1.1) with $\Phi^*(\mathbb{1}_d) = 2\mathrm{Re}(K)$. This is equivalent to L being representable as

$$L(\rho) = i[\rho, H] + \sum_j \mathcal{L}_j \rho \mathcal{L}_j^\dagger - \frac{1}{2}\{\mathcal{L}_j^\dagger \mathcal{L}_j, \rho\}\,, \tag{6.1.2}$$

where $H = H^\dagger \in \mathcal{M}_d$ is self-adjoint and $\{\mathcal{L}_j\}_j$ is a set of matrices in $\mathcal{M}_d$. Here, $\{\cdot,\cdot\}$ denotes the anti-commutator. We call such generators GKLS *or* Lindblad *generators.*

Remark 6.1.5. We shortly mention two infinite-dimensional generalizations of the results presented in this section. First, while we state Theorem 6.1.2 only in finite dimensions, it is more generally true that a uniformly continuous one-parameter semigroup of bounded linear maps on a Banach space admits a bounded generator [see, e.g., 160, Theorem I.3.7]. Indeed, even when we relax the continuity assumption from uniform to strong continuity, there still exists a closed and densely defined generator [see, e.g., 160, Theorem II.1.4].

Second, [16, Theorem 2] also contains infinite-dimensional analogues of Theorem 6.1.4. [16] characterized the bounded generators of uniformly continuous one-parameter semigroups of CPTP maps in the Schrödinger picture- In the Heisenberg picture, [16] established such a characterization under the additional assumption that all elements of the respective semigroup are normal CPU maps, i.e., CPU maps that are ultraweakly continuous.

As norm-continuous one-parameter semigroups are determined by their generators according to Theorem 6.1.2, it is tempting to see Theorem 6.1.4, with its characterization of these generators in the case of quantum dynamical semigroups, as providing a complete mathematical understanding of time-homogeneous Markovian quantum evolutions (at least in continuous time). However, there are interesting questions that Theorem 6.1.4 alone does not answer satisfactorily.
Concretely, consider the following task: Given a CPTP map $T \in \mathcal{CPTP}_d$, decide whether T is a member of some quantum dynamical semigroup. From the GKLS representation, we get an "answer" to this problem: $T \in \mathcal{CPTP}_d$ is a member of some quantum dynamical semigroup if and only if there exists a Lindblad generator L as in Eq. (6.1.2) such that $T = e^L$. While mathematically valid, this statement is practically useful only if there is an efficient procedure for deciding the existence of such a Lindblad generator. [161] showed that a rigorous version of this decision problem is hard in general, in fact NP-hard, as the system size grows. However, for a fixed system dimension d, we can determine whether a CPTP map $T \in \mathcal{CPTP}_d$ can be written approximately written as an exponential of a Lindblad generator efficiently in the desired accuracy [161, 162]. Therefore, for near-term quantum architectures, in which we can often effectively assume the system dimension to be a small constant, we can fit Lindblad generators to data gathered from quantum process tomography at a single point in time [162, 163].
The task of finding a best-fit Lindblad generator to a given CPTP map thus serves as an example of a computationally, mathematically, and physically fruitful problem in quantum Markovianity beyond the GKLS characterization. As a further natural question, we might ask for an understanding of Markovian quantum evolutions that need not be time-homogeneous, thereby leaving the framework of semigroups. Motivated by these and other questions, several different notions relating to quantum Markovianity have been introduced. Many of them are related to divisibility, originating from [10], or to infinitesimal deviations from complete positivity [164]. Others consider non-increasing distinguishability and (the lack of) quantum information backflow [165, 166]. Several review papers discuss the relations between these notions [166–169] and reiterating these connections is beyond the scope of this thesis. Rather, in Section 6.2, we explore the divisibility-based approach initiated by [10] in more detail.

6.2 Infinitesimal Markovian Divisible Quantum Channels

As discussed in the previous section, quantum dynamical semigroups model time-homogeneous Markovian quantum evolutions. We have seen that they correspond to solutions of so-called master equations, i.e., differential equations $\frac{\mathrm{d}}{\mathrm{d}t}T_t = LT_t$, $T_0 = \mathbb{1}_d$, with L a Lindblad generator. Now, a natural way of dropping the assumption of time-homogeneity is to consider solutions of differential equations $\frac{\mathrm{d}}{\mathrm{d}t}T_t = L_tT_t$, $T_0 = \mathbb{1}_d$, where L_t is a Lindblad generator that depends continuously on the time t. This serves as our motivation for the following definition:

Definition 6.2.1 (Infinitesimal Markovian Divisibility of Quantum Channels [10]). *Define the set* $\mathcal{I}_d \subset \mathcal{CPTP}_d$ *as*

$$\mathcal{I}_d \coloneqq \Big\{ T \in \mathcal{CPTP}_d \mid \forall \varepsilon > 0 \; \exists n \in \mathbb{N}, \; \textit{Lindblad generators } \{L_j\}_{j=1}^n \tag{6.2.1}$$

$$\textit{s.t. } (i) \left\| e^{L_j} - \mathbb{1}_d \right\| \leq \varepsilon \; \forall j \textit{ and } (ii) \prod_{j=1}^{n} e^{L_j} = T \Big\} . \tag{6.2.2}$$

We call the (operator norm-)closure $\overline{\mathcal{I}}_d$ *the set of* infinitesimal Markovian divisible quantum channels *(on qudits).*

Note that Definition 6.2.1 encompasses quantum dynamical semigroups in the following sense: If $T \in \mathcal{CPTP}_d$ is an element of a quantum dynamical semigroup, then in particular $T \in \mathcal{I}_d$.
Reference [10] provided some general insight into the structure of infinitesimal Markovian divisible quantum channels and completely characterized them in the qubit case. However, the only necessary criterion for a higher-dimensional quantum channel to be infinitesimal Markovian divisible observed in [10] was non-negativity of the determinant. This follows immediately from Definition 6.2.1 using continuity and multiplicativity of the determinant. In Core Article II [2], we complement this by showing that infinitesimal Markovian divisibility also implies upper bounds on the determinant in terms of products of smallest singular values. There, we consider notions of Markovian divisibility and infinitesimal Markovian divisibility for linear maps and general (compact and convex) sets of generators:

Definition 6.2.2 (Markovian Divisibility [2, Definition III.1]). *Let* $\mathcal{G} \subset \mathcal{B}(\mathbb{C}^d)$ *be a set of bounded linear maps, whose elements we call* generators. *We define the set*

$$\mathcal{D}_{\mathcal{G}} \coloneqq \{ T \in \mathcal{B}(\mathbb{C}^d) \mid \exists n \in \mathbb{N}, \; \textit{generators } \{G_i\}_{1 \leq i \leq n} \subset \mathcal{G} \; \textit{s.t.} \; \prod_{i=1}^{n} e^{G_i} = T \} . \tag{6.2.3}$$

We call the closure $\overline{\mathcal{D}}_{\mathcal{G}}$ *the set of linear maps that are* Markovian divisible w.r.t. $\mathcal{G}$.

Definition 6.2.3 (Infinitesimal Markovian Divisibility [2, Definition III.2]). *Let* $\mathcal{G} \subset \mathcal{B}(\mathbb{C}^d)$ *be a compact and convex set of bounded linear maps containing* $0 \in \mathcal{B}(\mathbb{C}^d)$. *We will again refer to elements of* $\mathcal{G}$ *as* generators. *We define the set*

$$\mathcal{I}_{\mathcal{G}} \coloneqq \{ T \in \mathcal{B}(\mathbb{C}^d) \mid \forall \varepsilon > 0 \; \exists n \in \mathbb{N}, \; \textit{generators } \{G_j\}_{1 \leq j \leq n} \subset \mathcal{G} \tag{6.2.4}$$

$$\textit{s.t. } (i) \left\| e^{G_j} - \mathbb{1}_d \right\| \leq \varepsilon \; \forall j \textit{ and } (ii) \prod_{j=1}^{n} e^{G_j} = T \} . \tag{6.2.5}$$

We call the closure $\overline{\mathcal{I}}_{\mathcal{G}}$ *the set of linear maps that are* infinitesimal Markovian divisible w.r.t. $\mathcal{G}$.

As explained in Remark III.3 of Core Article II [2], if we choose $\mathcal{G}$ to be the set of Lindblad generators on $\mathcal{B}(\mathbb{C}^d)$ with norm bounded by some strictly positive constant, then $\overline{\mathcal{I}}_d = \overline{\mathcal{D}}_{\mathcal{G}} = \overline{\mathcal{I}}_{\mathcal{G}}$. That is, we recover the notion from Definition 6.2.1.
In the first of the two main results of Core Article II [2], we show how to exploit majorization inequalities from matrix analysis together with Trotterization to prove the following result:

Theorem 6.2.4 ([2, Corollary IV.6]). *Let $\mathcal{G} \subset \mathcal{B}(\mathbb{C}^d)$ be a compact and convex set of bounded linear operators containing $0 \in \mathcal{B}(\mathbb{C}^d)$. Let $\tilde{\mathcal{G}} := \{\lambda G \mid \lambda \in [0,1],\ G \text{ an extreme point of } \mathcal{G}\} \subset \mathcal{G}$. Assume that every $\tilde{G} \in \tilde{\mathcal{G}}$ satisfies $\mathrm{tr}[\tilde{G} + \tilde{G}^*] - p \sum_{i=1}^{k} \lambda_i^{\uparrow}(\tilde{G} + \tilde{G}^*) \leq 0$. Let $T \in \overline{\mathcal{I}}_{\mathcal{G}}$. Then $0 \leq \det(T) \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T) \right)^p$.*

Theorem 6.2.4 allows us to derive upper bounds on the determinant of an infinitesimal Markovian divisible map from certain spectral properties of the real parts of admissible generators. A similar result also holds for elements of $\overline{\mathcal{D}}_{\mathcal{G}}$ [2, Theorem IV.5].
As our second central result in Core Article II [2], we prove that the real parts of Lindblad generators satisfy eigenvalue inequalities as needed in Theorem 6.2.4 when choosing the parameters $(p,k) = (d/2, 1)$ or $(p,k) = (1, \lfloor 2d - 2\sqrt{2d} + 1 \rfloor)$ [2, Lemmas IV.7 and IV.14]. Combining this with Theorem 6.2.4, we obtain the following necessary criteria for infinitesimal Markovian divisibility of quantum channels:

Corollary 6.2.5. *Let $T \in \overline{\mathcal{I}}_d$. Then we have*

$$0 \leq \det(T) \leq \min \left\{ \left(s_1^{\uparrow}(T) \right)^{\frac{d}{2}}, \prod_{i=1}^{\lfloor 2d - 2\sqrt{2d} + 1 \rfloor} s_i^{\uparrow}(T) \right\}. \tag{6.2.6}$$

Corollary 6.2.5 shows that singular value inequalities can serve as a tool for detecting "quantum Non-Markovianity" in the sense of a quantum channel not being infinitesimal Markovian divisible. They may also provide some guidance to an eventual characterization of infinitesimal Markovian divisible quantum channels beyond the qubit case.
In this section, we have discussed infinitesimal Markovian divisibility as an approach towards time-inhomogeneous quantum Markovianity. This generalizes the notion of quantum dynamical semigroups of quantum channels. The next section discusses a different extension of the framework of Section 6.1 by turning our attention from quantum channels to quantum superchannels.

6.3 Quantum Dynamical Semigroups of Quantum Superchannels

With the interest in higher-order quantum operations growing in recent years, and quantum Markovianity being an important topic in the study of "regular" quantum operations, it becomes natural to investigate Markovianity in higher-order quantum theory. As a first step in this direction, we go beyond quantum channels to quantum superchannels and consider continuous one-parameter semigroups thereof:

Definition 6.3.1 (Quantum Dynamical Semigroups of Superchannels [8]). *A family $(\hat{T}_t)_{t \geq 0}$ of linear maps $\hat{T}_t : \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B})) \to \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A}); \mathcal{B}(\mathbb{C}^{d_B}))$ forms a* quantum dynamical semigroup of superchannels *in the Schrödinger picture if*

(i) $\hat{T}_t$ is a quantum superchannel (according to Definition 2.3.1) for all $t \geq 0$,

(ii) $\hat{T}_{t+s} = \hat{T}_t \hat{T}_s$ for all $s, t \geq 0$,

(iii) $\hat{T}_0 = \mathrm{id}_{\mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A});\mathcal{B}(\mathbb{C}^{d_B}))}$, *and*

(iv) the map $\mathbb{R}_{\geq 0} \ni t \mapsto \hat{T}_t \in \mathcal{B}(\mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A});\mathcal{B}(\mathbb{C}^{d_B})))$ *is continuous.*

Again by Theorem 6.1.2, such quantum dynamical semigroups of superchannels are determined by their generators. The goal of Article VIII [8] was to characterize these generators, thereby proving a superchannel analogon of Theorem 6.1.4. From Section 2.3, we know that quantum superchannels are intimitately connected to semicausal, and therefore semilocalizable, CP maps. Thus, we aim to characterize the generators of continuous one-parameter semigroups of semicausal CP maps. It is easy to see that semicausality of the semigroup elements is equivalent to semicausality of the generator. Now, the main technical result of Article VIII [8] is the following normal form for semicausal Lindblad generators, which we state only in the Heisenberg picture and in finite dimensions for brevity:

Theorem 6.3.2 ([8, Theorem V.6]). *Let* $L \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}))$, $L(X) = \Phi(X) - K^\dagger X - XK$, *be a Lindblad generator, with* $\Phi \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}))$ *CP and* $K \in \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$. *Then,* L *is Heisenberg* $B \not\to A$ *semicausal if and only if there exists a Hilbert space* $\mathbb{C}^{d_E}$, *a unitary* $U \in \mathcal{B}(\mathbb{C}^{d_E} \otimes \mathbb{C}^{d_B};\mathbb{C}^{d_B} \otimes \mathbb{C}^{d_E})$, *a self-adjoint operator* $H_B \in \mathcal{B}(\mathbb{C}^{d_B})$, *and arbitrary operators* $A \in \mathcal{B}(\mathbb{C}^{d_A};\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_E})$, $B \in \mathcal{B}(\mathbb{C}^{d_B};\mathbb{C}^{d_B} \otimes \mathbb{C}^{d_E})$ *and* $K_A \in \mathcal{B}(\mathbb{C}^{d_A})$, *such that*

$$\Phi(X) = V^\dagger (X \otimes \mathbb{1}_E) V\,, \text{ with } V = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B) + (\mathbb{1}_A \otimes B)\,, \tag{6.3.1}$$

$$K = (\mathbb{1}_A \otimes B^\dagger U)(A \otimes \mathbb{1}_B) + \frac{1}{2}\mathbb{1}_A \otimes B^\dagger B + K_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes iH_B\,. \tag{6.3.2}$$

Using Theorem 2.3.5, we translate this to a complete characterization of the generators of quantum dynamical semigroups of superchannels:

Theorem 6.3.3 ([8, Theorem V.17]). *A linear map* $\hat{L} : \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A});\mathcal{B}(\mathbb{C}^{d_B})) \to \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A});\mathcal{B}(\mathbb{C}^{d_B}))$ *generates a semigroup of superchannels if and only if there exists there exists a Hilbert space* $\mathbb{C}^{d_E}$, *a state* $\sigma \in \mathcal{S}(\mathbb{C}^{d_E})$, *a unitary* $U \in \mathcal{B}(\mathbb{C}^{d_B} \otimes \mathbb{C}^{d_E})$, *a self-adjoint operator* $H_B \in \mathcal{B}(\mathbb{C}^{d_B})$, *and arbitrary operators* $A \in \mathcal{B}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_E})$, $B \in \mathcal{B}(\mathbb{C}^{d_B} \otimes \mathbb{C}^{d_E})$ *and* $K_A \in \mathcal{B}(\mathbb{C}^{d_A})$, *satisfying that* $\mathrm{tr}_\sigma\left[A^\dagger A\right] = K_A + K_A^\dagger$ *and that* $\hat{L}$ *acts on* $T \in \mathcal{B}(\mathcal{B}(\mathbb{C}^{d_A});\mathcal{B}(\mathbb{C}^{d_B}))$ *as*

$$\hat{L}(T) = \hat{\Phi}(T) - \hat{\kappa}_L(T) - \hat{\kappa}_R(T)\,, \tag{6.3.3}$$

with

$$\begin{aligned}\hat{\Phi}(T)(\rho) = {}& \mathrm{tr}_E\left[U\,(T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma)A^\dagger)\,U^\dagger\right] + \mathrm{tr}_E\left[B\,(T \otimes \mathrm{id}_E)((\rho \otimes \sigma)A^\dagger)\,U^\dagger\right] \\ & + \mathrm{tr}_E\left[U\,(T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma))\,B^\dagger\right] + \mathrm{tr}_E\left[B\,(T \otimes \mathrm{id}_E)((\rho \otimes \sigma))\,B^\dagger\right]\,, \end{aligned} \tag{6.3.4}$$

$$\begin{aligned}\hat{\kappa}_L(T)(\rho) = {}& \mathrm{tr}_E\left[B^\dagger U\,(T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma))\right] + \frac{1}{2}\mathrm{tr}_E\left[B^\dagger B(T \otimes \mathrm{id}_E)(\rho \otimes \sigma)\right] \\ & + T(K_A\,\rho) + iH_B\,T(\rho)\,, \end{aligned} \tag{6.3.5}$$

$$\begin{aligned}\hat{\kappa}_R(T)(\rho) = {}& \mathrm{tr}_E\left[(T \otimes \mathrm{id}_E)((\rho \otimes \sigma)A^\dagger)\,U^\dagger B\right] + \frac{1}{2}\mathrm{tr}_E\left[(T \otimes \mathrm{id}_E)(\rho \otimes \sigma)B^\dagger B\right] \\ & + T(\rho\,K_A^\dagger) - T(\rho)\,iH_B\,. \end{aligned} \tag{6.3.6}$$

Bibliography

[1] Matthias C. Caro and Ishaun Datta. 'Pseudo-dimension of quantum circuits'. In: *Quantum Machine Intelligence* 2, 14 (2020). DOI: 10.1007/s42484-020-00027-5 (cit. on pp. xi, 3, 5, 17, 41–43).

[2] Matthias C. Caro and Benedikt R. Graswald. 'Necessary criteria for Markovian divisibility of linear maps'. In: *Journal of Mathematical Physics* 62.4, 042203 (2021). DOI: 10.1063/5.0031760 (cit. on pp. xi, 3, 17, 61, 62).

[3] Matthias C. Caro. 'Binary Classification with Classical Instances and Quantum Labels'. In: *Quantum Machine Intelligence* 3, 18 (2021). DOI: 10.1007/s42484-021-00043-z (cit. on pp. xi, 4, 17, 52–55).

[4] Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. 'Encoding-dependent generalization bounds for parametrized quantum circuits'. In: *Quantum* 5, 582 (2021). DOI: 10.22331/q-2021-11-17-582 (cit. on pp. xi, 4, 5, 17, 44–46).

[5] Matthias C. Caro. 'Quantum learning Boolean linear functions w.r.t. product distributions'. In: *Quantum Information Processing* 19, 172 (2020). DOI: 10.1007/s11128-020-02661-1 (cit. on pp. xi, 5, 17, 48, 51, 52).

[6] Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. 'Generalization in quantum machine learning from few training data'. In: *Nature Communications* 13, 4919 (2022). DOI: 10.1038/s41467-022-32550-3 (cit. on pp. xi, 5, 17, 41–44).

[7] Matthias C. Caro. *Undecidability of Learnability*. Version 2. Aug. 1, 2021. arXiv: 2106.01382 [cs.CC] (cit. on pp. xii, 6, 32, 33).

[8] Markus Hasenöhrl and Matthias C Caro. 'Quantum and classical dynamical semigroups of superchannels and semicausal channels'. In: *Journal of Mathematical Physics* 63.7 (2022), p. 072204. DOI: 10.1063/5.0070635 (cit. on pp. xii, 6, 17, 62, 63).

[9] Hugh E. Warren. 'Lower bounds for approximation by nonlinear manifolds'. In: *Transactions of the American Mathematical Society* 133.1 (1968), pp. 167–178. DOI: 10.1090/S0002-9947-1968-0226281-1 (cit. on pp. 3, 41).

[10] Michael M. Wolf and J. Ignacio Cirac. 'Dividing quantum channels'. In: *Communications in Mathematical Physics* 279.1 (2008), pp. 147–168. DOI: 10.1007/s00220-008-0411-y (cit. on pp. 3, 60, 61, 99).

[11] Andrew W. Cross, Graeme Smith, and John A. Smolin. 'Quantum learning robust against noise'. In: *Phys. Rev. A* 92 (1 July 2015), p. 012327. DOI: 10.1103/PhysRevA.92.012327 (cit. on pp. 5, 48, 50, 51).

[12] Alex B. Grilo, Iordanis Kerenidis, and Timo Zijlstra. 'Learning-with-errors problem is easy with quantum samples'. In: *Phys. Rev. A* 99 (3 Mar. 2019), p. 032314. DOI: 10.1103/PhysRevA.99.032314 (cit. on pp. 5, 48, 50, 51).

[13] Ethan Bernstein and Umesh Vazirani. 'Quantum Complexity Theory'. In: *SIAM J. Comput.* 26.5 (Oct. 1997), pp. 1411–1473. ISSN: 0097-5397. DOI: 10.1137/S0097539796300921 (cit. on pp. 5, 50, 51).

[14] Varun Kanade, Andrea Rocchetto, and Simone Severini. 'Learning DNFs under product distributions via μ-biased quantum Fourier sampling'. In: *Quantum Information & Computation* 19.15&16 (2019), pp. 1261–1278. DOI: 10.26421/QIC19.15-16 (cit. on pp. 5, 48, 50).

[15] Vittorio Gorini, Andrzej Kossakowski, and E. C. G. Sudarshan. 'Completely positive dynamical semigroups of N-level systems'. In: *Journal of Mathematical Physics* 17.5 (1976), p. 821. ISSN: 00222488. DOI: 10.1063/1.522979 (cit. on pp. 6, 59).

[16] Göran Lindblad. 'On the generators of quantum dynamical semigroups'. In: *Communications in Mathematical Physics* 48.2 (1976), pp. 119–130. DOI: 10.1007/BF01608499 (cit. on pp. 6, 59, 60).

[17] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000 (cit. on pp. 9, 12).

[18] Teiko Heinosaari and Mário Ziman. *The mathematical language of quantum theory: from uncertainty to entanglement*. Cambridge University Press, 2011 (cit. on p. 9).

[19] John Watrous. *The theory of quantum information*. Cambridge University Press, 2018 (cit. on p. 9).

[20] Michael M. Wolf. *Quantum channels & operations: Guided tour*. July 5, 2012. URL: https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MichaelWolf/QChannelLecture.pdf (cit. on pp. 9, 14, 15).

[21] John Preskill. *Quantum Computation (Lecture Notes)*. Nov. 2020. URL: http://theory.caltech.edu/~preskill/ph219/index.html#lecture (cit. on pp. 9, 11).

[22] Ronald de Wolf. *Quantum Computing: Lecture Notes*. Version 2. Jan. 20, 2021. arXiv: arXiv:1907.09415 [quant-ph] (cit. on p. 9).

[23] Christopher A. Fuchs and Jeroen Van De Graaf. 'Cryptographic distinguishability measures for quantum-mechanical states'. In: *IEEE Transactions on Information Theory* 45.4 (1999), pp. 1216–1227. DOI: 10.1109/18.761271 (cit. on p. 12).

[24] Andrzej Jamiołkowski. 'Linear transformations which preserve trace and positive semidefiniteness of operators'. In: *Reports on Mathematical Physics* 3.4 (1972), pp. 275–278. DOI: 10.1016/0034-4877(72)90011-0 (cit. on p. 14).

[25] Man-Duen Choi. 'Completely positive linear maps on complex matrices'. In: *Linear algebra and its applications* 10.3 (1975), pp. 285–290. DOI: 10.1016/0024-3795(75)90075-0 (cit. on p. 14).

[26] Karl Kraus. *States, Effects, and Operations Fundamental Notions of Quantum Theory*. Springer, 1983 (cit. on p. 14).

[27] W. Forrest Stinespring. 'Positive functions on C*-algebras'. In: *Proceedings of the American Mathematical Society* 6.2 (1955), pp. 211–216. DOI: 10.2307/2032342 (cit. on p. 14).

[28] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. 'Transforming quantum operations: Quantum supermaps'. In: *EPL (Europhysics Letters)* 83.3 (2008), p. 30004. DOI: 10.1209/0295-5075/83/30004 (cit. on pp. 16, 17).

[29] David Beckman, Daniel Gottesman, Michael A. Nielsen, and John Preskill. 'Causal and localizable quantum operations'. In: *Physical Review A* 64.5 (2001), p. 052309. DOI: 10.1103/PhysRevA.64.052309 (cit. on p. 16).

[30] Tilo Eggeling, Dirk Schlingemann, and Reinhard F. Werner. 'Semicausal operations are semilocalizable'. In: *EPL (Europhysics Letters)* 57.6 (2002), p. 782. DOI: 10.1209/epl/i2002-00579-4 (cit. on pp. 16, 17).

[31] Marco Piani, Michal Horodecki, Pawel Horodecki, and Ryszard Horodecki. 'Properties of quantum nonsignaling boxes'. In: *Physical Review A* 74.1 (2006), p. 012305. DOI: 10.1103/PhysRevA.74.012305 (cit. on p. 17).

[32] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994 (cit. on p. 19).

[33] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999 (cit. on pp. 19, 29, 30, 40).

[34] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018 (cit. on pp. 19, 23, 27).

[35] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2019 (cit. on p. 19).

[36] Michael M. Wolf. *Mathematical Foundations of Supervised Learning (growing lecture notes)*. July 5, 2021 (cit. on pp. 19, 23, 27, 29, 30, 39, 42).

[37] David Haussler. 'Decision theoretic generalizations of the PAC model for neural net and other learning applications'. In: *Information and Computation* 100.1 (1992), pp. 78–150. DOI: 10.1016/0890-5401(92)90010-D (cit. on pp. 20, 21).

[38] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. 'Toward efficient agnostic learning'. In: *Machine Learning* 17.2-3 (1994), pp. 115–141. DOI: 10.1007/BF00993468 (cit. on pp. 20, 21).

[39] Vladimir N. Vapnik and Alexei Ya. Chervonenkis. 'On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities'. In: *Theory of Probability & Its Applications* 16.2 (1971), pp. 264–280. DOI: 10.1137/1116025 (cit. on pp. 21, 23).

[40] Leslie G. Valiant. 'A Theory of the Learnable'. In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142. ISSN: 00010782. DOI: 10.1145/1968.1972 (cit. on p. 21).

[41] David A. McAllester. 'Some PAC-Bayesian Theorems'. In: *Machine Learning* 37.3 (1999), pp. 355–363. DOI: 10.1023/A:1007618624809 (cit. on pp. 23, 31).

[42] Dana Angluin. 'Queries and concept learning'. In: *Machine learning* 2.4 (1988), pp. 319–342. DOI: 10.1023/A:1022821128753 (cit. on p. 23).

[43] Michael J. Kearns. 'Efficient noise-tolerant learning from statistical queries'. In: *Journal of the ACM (JACM)* 45.6 (1998), pp. 983–1006. DOI: 10.1145/293347.293351 (cit. on p. 23).

[44] Nick Littlestone. 'Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm'. In: *Machine Learning* 2.4 (1988), pp. 285–318. ISSN: 1573-0565. DOI: 10.1023/A:1022869011914 (cit. on pp. 23, 33).

[45] Martin Zinkevich. 'Online convex programming and generalized infinitesimal gradient ascent'. In: *Proceedings of the 20th International Conference on Machine Learning*. 2003, pp. 928–936. URL: https://www.aaai.org/Library/ICML/2003/icml03-120.php (cit. on p. 23).

[46] Sally A. Goldman and Michael J. Kearns. 'On the Complexity of Teaching'. In: *Journal of Computer and System Sciences* 50.1 (1995), pp. 20–31. ISSN: 00220000. DOI: 10.1006/jcss.1995.1003 (cit. on pp. 23, 33).

[47] Frank J. Balbach. 'Measuring teachability using variants of the teaching dimension'. In: *Theoretical Computer Science* 397.1-3 (2008), pp. 94–113. ISSN: 03043975. DOI: 10.1016/j.tcs.2008.02.025 (cit. on p. 23).

[48] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. 'Models of Cooperative Teaching and Learning'. In: *Journal of Machine Learning Research* 12.11 (2011), pp. 349–384. URL: http://jmlr.org/papers/v12/zilles11a.html (cit. on p. 23).

[49] Vladimir Vapnik and Alexey Chervonenkis. *Theory of pattern recognition*. Nauka, Moscow, 1974. (in Russian) (cit. on p. 24).

[50] Michel Talagrand. 'Sharper bounds for Gaussian and empirical processes'. In: *The Annals of Probability* (1994), pp. 28–76. DOI: 10.1214/aop/1176988847 (cit. on p. 24).

[51] Steve Hanneke. 'The Optimal Sample Complexity of PAC Learning'. In: *Journal of Machine Learning Research* 17.38 (2016), pp. 1–15. URL: http://jmlr.org/papers/v17/15-389.html (cit. on pp. 24, 55).

[52] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 'Learnability and the Vapnik-Chervonenkis dimension'. In: *Journal of the ACM (JACM)* 36.4 (1989), pp. 929–965. DOI: 10.1145/76359.76371 (cit. on p. 24).

[53] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. 'A general lower bound on the number of examples needed for learning'. In: *Information and Computation* 82.3 (1989), pp. 247–261. DOI: 10.1016/0890-5401(89)90002-3 (cit. on p. 24).

[54] Hans Ulrich Simon. 'General Bounds on the Number of Examples Needed for Learning Probabilistic Concepts'. In: *Journal of Computer and System Sciences* 52.2 (1996), pp. 239–254. DOI: 10.1006/jcss.1996.0019 (cit. on p. 24).

[55] David Pollard. *Convergence of stochastic processes*. Springer, 1984 (cit. on p. 24).

[56] Michael J. Kearns and Robert E. Schapire. 'Efficient distribution-free learning of probabilistic concepts'. In: *Journal of Computer and System Sciences* 48.3 (1994), pp. 464–497. DOI: 10.1016/S0022-0000(05)80062-5 (cit. on p. 25).

[57] Martin Anthony and Peter L. Bartlett. 'Function learning from interpolation'. In: *Combinatorics, Probability and Computing* 9.3 (2000), pp. 213–225. DOI: 10.1017/S0963548300004247 (cit. on pp. 25, 42, 43).

[58] Evarist Giné and Joel Zinn. 'Some Limit Theorems for Empirical Processes'. In: *The Annals of Probability* 12.4 (1984), pp. 929–989. DOI: 10.1214/aop/1176993138 (cit. on p. 25).

[59] Colin McDiarmid. 'On the method of bounded differences'. In: *Surveys in combinatorics, 1989 (Norwich, 1989)*. Vol. 141. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, 1989, pp. 148–188. DOI: 10.1017/CBO9781107359949.008 (cit. on p. 26).

[60] Peter L. Bartlett and Shahar Mendelson. 'Rademacher and Gaussian Complexities: Risk Bounds and Structural Results'. In: *Journal of Machine Learning Research* 3 (Nov. 2002), pp. 463–482. URL: https://jmlr.org/papers/v3/bartlett02a.html (cit. on pp. 26, 27).

[61] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 1991 (cit. on pp. 26, 27).

[62] Vladimir Koltchinskii and Dmitriy Panchenko. 'Rademacher Processes and Bounding the Risk of Function Learning'. In: *Giné E., Mason D.M., Wellner J.A. (eds) High Dimensional Probability II*. Progress in Probability. Springer Science & Business Media, 2000, pp. 443–457. DOI: 10.1007/978-1-4612-1358-1_29 (cit. on p. 27).

[63] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018. DOI: 10.1017/9781108231596 (cit. on pp. 28–31, 42).

[64] Richard M. Dudley. 'The sizes of compact subsets of Hilbert space and continuity of Gaussian processes'. In: *Journal of Functional Analysis* 1.3 (1967), pp. 290–330. DOI: 10.1016/0022-1236(67)90017-1 (cit. on p. 28).

[65] Richard M. Dudley. 'A course on empirical processes'. In: *Ecole d'été de Probabilités de Saint-Flour XII-1982*. Springer, 1984, pp. 1–142. DOI: 10.1007/BFb0099432 (cit. on p. 28).

[66] Shahar Mendelson and Roman Vershynin. 'Entropy and the combinatorial dimension'. In: *Inventiones mathematicae* 152.1 (2003), pp. 37–55. DOI: 10.1007/s00222-002-0266-3 (cit. on p. 30).

[67] Mathukumalli Vidyasagar. *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer, 2003 (cit. on p. 30).

[68] Richard M. Dudley. *Uniform central limit theorems*. Cambridge University Press, 1999 (cit. on p. 30).

[69] Nick Littlestone and Manfred Warmuth. *Relating data compression and learnability*. 1986. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.7589&rep=rep1&type=pdf. Technical report, University of California Santa Cruz (cit. on p. 31).

[70] Olivier Bousquet and André Elisseeff. 'Stability and Generalization'. In: *The Journal of Machine Learning Research* 2 (Mar. 2002), pp. 499–526. URL: https://jmlr.org/papers/v2/bousquet02a.html (cit. on p. 31).

[71] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 'Calibrating noise to sensitivity in private data analysis'. In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284. DOI: 10.1007/11681878_14 (cit. on p. 31).

[72] Richard H. Lathrop. 'On the learnability of the uncomputable'. In: *Proc. 13th International Conference on Machine Learning*. Morgan Kaufmann, 1996, pp. 302–309. URL: https://www.ics.uci.edu/~rickl/publications/1996-icml.pdf (cit. on p. 32).

[73] Marcus Schaefer. 'Deciding the Vapnik–Červonenkis Dimension is Σ_3^p-complete'. In: *Journal of Computer and System Sciences* 58.1 (1999), pp. 177–182. DOI: 10.1006/jcss.1998.1602 (cit. on pp. 32, 33).

[74] Kino Zhao. *A statistical learning approach to a problem of induction*. Dec. 8, 2018. URL: http://philsci-archive.pitt.edu/15422/ (cit. on pp. 32, 33).

[75] Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. 'Learnability can be undecidable'. In: *Nature Machine Intelligence* 1.1 (2019), pp. 44–48. ISSN: 2522-5839. DOI: 10.1038/s42256-018-0002-3 (cit. on pp. 32, 33).

[76] Sushant Agarwal, Nivasini Ananthakrishnan, Shai Ben-David, Tosca Lechner, and Ruth Urner. 'On Learnability wih Computable Learners'. In: *Algorithmic Learning Theory* (2020), pp. 48–60. ISSN: 1938-7228. URL: http://proceedings.mlr.press/v117/agarwal20b.html (cit. on pp. 32, 33).

[77] Kurt Gödel. 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I'. In: *Monatshefte für Mathematik und Physik* 38.1 (1931), pp. 173–198. ISSN: 1436-5081. DOI: 10.1007/BF01700692 (cit. on p. 32).

[78] Alan M. Turing. 'On Computable Numbers, with an Application to the Entscheidungsproblem'. In: *Proceedings of the London Mathematical Society* s2-42.1 (1937), pp. 230–265. ISSN: 0024-6115. DOI: 10.1112/plms/s2-42.1.230 (cit. on p. 32).

[79] Srinivasan Arunachalam and Ronald de Wolf. 'Optimal Quantum Sample Complexity of Learning Algorithms'. In: *Journal of Machine Learning Research* 19.71 (2018), pp. 1–36. URL: http://jmlr.org/papers/v19/18-195.html (cit. on pp. 33, 48, 56).

[80] Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. *A Theory of Universal Learning*. Version 1. Nov. 9, 2020. arXiv: 2011.04483 [cs.LG] (cit. on p. 33).

[81] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. 'Quantum circuit learning'. In: *Physical Review A* 98.3 (2018), p. 032309. DOI: 10.1103/PhysRevA.98.032309 (cit. on p. 35).

[82] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. 'Evaluating analytic gradients on quantum hardware'. In: *Physical Review A* 99.3 (2019), p. 032331. DOI: 10.1103/PhysRevA.99.032331 (cit. on p. 35).

[83] John Preskill. 'Quantum computing in the NISQ era and beyond'. In: *Quantum* 2 (2018), p. 79. DOI: 10.22331/q-2018-08-06-79 (cit. on p. 35).

[84] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L O'brien. 'A variational eigenvalue solver on a photonic quantum processor'. In: *Nature communications* 5.1 (2014), pp. 1–7. DOI: 10.1038/ncomms5213 (cit. on p. 35).

[85] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. *A Quantum Approximate Optimization Algorithm*. Version 1. Nov. 14, 2014. arXiv: 1411.4028 [quant-ph] (cit. on p. 35).

[86] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. 'Variational quantum algorithms'. In: *Nature Reviews Physics* (2021), pp. 1–20. DOI: 10.1038/s42254-021-00348-9 (cit. on p. 36).

[87] Maria Schuld and Nathan Killoran. 'Quantum machine learning in feature Hilbert spaces'. In: *Physical review letters* 122.4 (2019), p. 040504. DOI: 10.1103/PhysRevLett.122.040504 (cit. on pp. 36, 39).

[88] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. 'Supervised learning with quantum-enhanced feature spaces'. In: *Nature* 567.7747 (2019), pp. 209–212. DOI: 10.1038/s41586-019-0980-2 (cit. on p. 36).

[89] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. 'Parameterized quantum circuits as machine learning models'. In: *Quantum Science and Technology* 4.4 (2019), p. 043001. DOI: 10.1088/2058-9565/ab4eb5 (cit. on p. 36).

[90] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. 'Barren plateaus in quantum neural network training landscapes'. In: *Nature communications* 9.1 (2018), pp. 1–6. DOI: 10.1038/s41467-018-07090-4 (cit. on p. 36).

[91] Arthur Pesah, M. Cerezo, Samson Wang, Tyler Volkoff, Andrew T. Sornborger, and Patrick J. Coles. *Absence of Barren Plateaus in Quantum Convolutional Neural Networks*. Version 1. Nov. 5, 2020. arXiv: 2011.02966 [quant-ph] (cit. on p. 36).

[92] Carlos Ortiz Marrero, Mária Kieferová, and Nathan Wiebe. *Entanglement Induced Barren Plateaus*. Version 1. Mar. 10, 2021. arXiv: 2010.15968 [quant-ph] (cit. on p. 36).

[93] M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. 'Cost function dependent barren plateaus in shallow parametrized quantum circuits'. In: *Nature Communications* 12.1 (Mar. 19, 2021), p. 1791. DOI: 10.1038/s41467-021-21728-w (cit. on p. 36).

[94] Alexey V. Uvarov and Jacob D. Biamonte. 'On barren plateaus and cost function locality in variational quantum algorithms'. In: *Journal of Physics A: Mathematical and Theoretical* 54.24 (2021), p. 245301. DOI: 10.1088/1751-8121/abfac7 (cit. on p. 36).

[95] Yang Qian, Xinbiao Wang, Yuxuan Du, Xingyao Wu, and Dacheng Tao. *The dilemma of quantum neural networks*. Version 1. June 9, 2021. arXiv: 2106.04975 [quant-ph] (cit. on p. 36).

[96] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. 'The power of quantum neural networks'. In: *Nature Computational Science* 1.6 (2021), pp. 403–409. DOI: 10.1038/s43588-021-00084-1 (cit. on pp. 36, 40).

[97] Chih-Chieh Chen, Masaya Watabe, Kodai Shiba, Masaru Sogabe, Katsuyoshi Sakamoto, and Tomah Sogabe. 'On the Expressibility and Overfitting of Quantum Circuit Learning'. In: *ACM Transactions on Quantum Computing* 2.2 (2021), pp. 1–24. DOI: 10.1145/3466797 (cit. on p. 36).

[98] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I. Latorre. 'Data re-uploading for a universal quantum classifier'. In: *Quantum* 4 (2020), p. 226. DOI: 10.22331/q-2020-02-06-226 (cit. on p. 37).

[99] Bernhard Schölkopf and Alex J. Smola. *Learning with Kernels*. MIT Press, 2002 (cit. on p. 39).

[100] George Kimeldorf and Grace Wahba. 'Some results on Tchebycheffian spline functions'. In: *Journal of Mathematical Analysis and Applications* 33.1 (1971), pp. 82–95. DOI: 10.1016/0022-247X(71)90184-3 (cit. on p. 39).

[101] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. 'A Generalized Representer Theorem'. In: *International Conference on Computational Learning Theory*. Springer. 2001, pp. 416–426. DOI: 10.1007/3-540-44581-1_27 (cit. on p. 39).

[102] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. 'Power of data in quantum machine learning'. In: *Nature Communications* 12.1 (2021), pp. 1–9. DOI: 10.1038/s41467-021-22539-9 (cit. on p. 39).

[103] Maria Schuld. *Supervised quantum machine learning models are kernel methods*. Version 2. Apr. 17, 2021. arXiv: 2101.11020 [quant-ph] (cit. on p. 39).

[104] Sofiene Jerbi, Lukas J. Fiderer, Hendrik Poulsen Nautrup, Jonas M. Kübler, Hans J. Briegel, and Vedran Dunjko. *Quantum machine learning beyond kernel methods*. Version 1. Oct. 25, 2021. arXiv: 2110.13162 [quant-ph] (cit. on p. 39).

[105] Casper Gyurik, Dyon van Vreumingen, and Vedran Dunjko. *Structural risk minimization for quantum linear classifiers*. Version 1. May 12, 2021. arXiv: 2105.05566 [quant-ph] (cit. on p. 40).

[106] Amira Abbas, David Sutter, Alessio Figalli, and Stefan Woerner. *Effective dimension of machine learning models*. Version 1. Dec. 9, 2021. arXiv: 2112.04807 [cs.LG] (cit. on p. 40).

[107] Leonardo Banchi, Jason Pereira, and Stefano Pirandola. 'Generalization in Quantum Machine Learning: A Quantum Information Standpoint'. In: *PRX Quantum* 2.4 (2021), p. 040321. DOI: 10.1103/PRXQuantum.2.040321 (cit. on p. 40).

[108] Marek Karpinski and Angus Macintyre. 'Polynomial Bounds for VC Dimension of Sigmoidal and General Pfaffian Neural Networks'. In: *Journal of Computer and System Sciences* 54.1 (1997), pp. 169–176. DOI: 10.1006/jcss.1997.1477 (cit. on p. 40).

[109] Paul W. Goldberg and Mark R. Jerrum. 'Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers'. In: *Machine Learning* 18.2-3 (1995), pp. 131–148. DOI: 10.1007/BF00993408 (cit. on p. 41).

[110] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. 'Sample-optimal tomography of quantum states'. In: *IEEE Transactions on Information Theory* 63.9 (2017), pp. 5628–5641. DOI: 10.1109/TIT.2017.2719044 (cit. on p. 43).

[111] Hao-Chung Cheng, Min-Hsiu Hsieh, and Ping-Cheng Yeh. 'The learnability of unknown quantum measurements'. In: *Quantum Inf. Comput.* 16.7&8 (2016), pp. 615–656. DOI: 10.26421/QIC16.7-8-4 (cit. on p. 43).

[112] Andrea Rocchetto. 'Stabiliser States Are Efficiently PAC-Learnable'. In: *Quantum Info. Comput.* 18.7–8 (June 2018), pp. 541–552. DOI: 10.26421/QIC18.7-8-1 (cit. on p. 43).

[113] Scott Aaronson. 'Shadow Tomography of Quantum States'. In: *SIAM Journal on Computing* 49.5 (2019), STOC18–368–STOC18–394. DOI: 10.1145/3188745.3188802 (cit. on p. 43).

[114] Scott Aaronson, Xinyi Chen, Elad Hazan, Satyen Kale, and Ashwin Nayak. 'Online learning of quantum states'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (Dec. 2019), p. 124019. DOI: 10.1088/1742-5468/ab3988 (cit. on p. 43).

[115] Scott Aaronson and Guy N. Rothblum. 'Gentle measurement of quantum states and differential privacy'. In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 322–333. DOI: 10.1145/3313276.3316378 (cit. on p. 43).

[116] Mithuna Yoganathan. *A condition under which classical simulability implies efficient state learnability*. Version 1. July 18, 2019. arXiv: 1907.08163 [quant-ph] (cit. on p. 43).

[117] Costin Bădescu and Ryan O'Donnell. *Improved quantum data analysis*. Version 1. Nov. 22, 2020. arXiv: 2011.10908 [quant-ph] (cit. on p. 43).

[118] Hsin-Yuan Huang, Richard Kueng, and John Preskill. 'Predicting many properties of a quantum system from very few measurements'. In: *Nature Physics* 16.10 (2020), pp. 1050–1057. DOI: 10.1038/s41567-020-0932-7 (cit. on p. 43).

[119] Srinivasan Arunachalam, Yihui Quek, and John Smolin. *Private learning implies quantum stability*. Version 1. Feb. 14, 2021. arXiv: 2102.07171 [quant-ph] (cit. on p. 43).

[120] Anurag Anshu, Srinivasan Arunachalam, Tomotaka Kuwahara, and Mehdi Soleimanifar. 'Sample-efficient learning of interacting quantum systems'. In: *Nature Physics* (2021), pp. 1–5. DOI: 10.1038/s41567-021-01232-0 (cit. on p. 43).

[121] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill. *Provably efficient machine learning for quantum many-body problems*. Version 2. July 18, 2021. arXiv: 2106.12627 [quant-ph] (cit. on p. 43).

[122] Cambyse Rouzé and Daniel Stilck França. *Learning quantum many-body systems from a few copies*. Version 1. July 7, 2021. arXiv: 2107.03333 [quant-ph] (cit. on p. 43).

[123] Jeongwan Haah, Robin Kothari, and Ewin Tang. *Optimal learning of quantum Hamiltonians from high-temperature Gibbs states*. Version 1. Aug. 10, 2021. arXiv: 2108.04842 [quant-ph] (cit. on p. 43).

[124] Scott Aaronson. 'The learnability of quantum states'. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 463.2088 (2007), pp. 3089–3114. DOI: 10.1098/rspa.2007.0113 (cit. on p. 43).

[125] Claudiu Marius Popescu. 'Learning bounds for quantum circuits in the agnostic setting'. In: *Quantum Information Processing* 20, 286 (2021). DOI: 10.1007/s11128-021-03225-7 (cit. on p. 43).

[126] Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. *An efficient measure for the expressivity of variational quantum algorithms*. Version 1. Apr. 20, 2021. arXiv: 2104.09961 [quant-ph] (cit. on pp. 43, 44).

[127] Francisco Javier Gil Vidal and Dirk Oliver Theis. 'Input Redundancy for Parameterized Quantum Circuits'. In: *Frontiers in Physics* 8 (2020), p. 297. DOI: 10.3389/fphy.2020.00297 (cit. on p. 44).

[128] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. 'Effect of data encoding on the expressive power of variational quantum-machine-learning models'. In: *Physical Review A* 103.3 (2021), p. 032430. DOI: 10.1103/PhysRevA.103.032430 (cit. on p. 44).

[129] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta. 'Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets'. In: *Nature* 549.7671 (2017), pp. 242–246. DOI: 10.1038/nature23879 (cit. on p. 44).

[130] Rocco A. Servedio and Steven J. Gortler. 'Equivalences and Separations Between Quantum and Classical Learnability'. In: *SIAM Journal on Computing* 33.5 (2004), pp. 1067–1092. DOI: 10.1137/S0097539704412910 (cit. on p. 47).

[131] Ashley Montanaro. 'The quantum query complexity of learning multilinear polynomials'. In: *Information Processing Letters* 112.11 (2012), pp. 438–442. DOI: 10.1016/j.ipl.2012.03.002 (cit. on p. 47).

[132] Srinivasan Arunachalam, Sourav Chakraborty, Troy Lee, Manaswi Paraashar, and Ronald de Wolf. 'Two New Results About Quantum Exact Learning'. In: *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Ed. by Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi. Vol. 132. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 16:1–16:15. ISBN: 978-3-95977-109-2. DOI: 10.4230/LIPIcs.ICALP.2019.16 (cit. on pp. 47, 48, 50).

[133] Srinivasan Arunachalam, Alex B. Grilo, and Henry Yuen. *Quantum statistical query learning*. Version 2. Nov. 24, 2020. arXiv: 2002.08240 [quant-ph] (cit. on p. 47).

[134] Nader H. Bshouty and Jeffrey C. Jackson. 'Learning DNF over the Uniform Distribution Using a Quantum Example Oracle'. In: *SIAM Journal on Computing* 28.3 (1998), pp. 1136–1153. ISSN: 0097-5397. DOI: 10.1137/S0097539795293123 (cit. on pp. 47, 48, 50).

[135] Srinivasan Arunachalam and Ronald de Wolf. 'Guest Column: A Survey of Quantum Learning Theory'. In: *SIGACT News* 48 (2017). DOI: 10.1145/3106700.3106710 (cit. on p. 47).

[136] Scott Aaronson. 'Read the fine print'. In: *Nature Physics* 11.4 (2015), pp. 291–293. DOI: 10.1038/nphys3272 (cit. on p. 48).

[137] Carlo Ciliberto, Mark Herbster, Alessandro Davide Ialongo, Massimiliano Pontil, Andrea Rocchetto, Simone Severini, and Leonard Wossnig. 'Quantum machine learning: a classical perspective'. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 474.2209 (2018), p. 20170551. DOI: 10.1098/rspa.2017.0551 (cit. on p. 48).

[138] Alp Atıcı and Rocco A. Servedio. 'Improved Bounds on Quantum Learning Algorithms'. In: *Quantum Information Processing* 4.5 (2005), pp. 355–386. DOI: 10.1007/s11128-005-0001-2 (cit. on p. 48).

[139] Chi Zhang. 'An improved lower bound on query complexity for quantum PAC learning'. In: *Information Processing Letters* 111.1 (2010), pp. 40–45. DOI: 10.1016/j.ipl.2010.10.007 (cit. on p. 48).

[140] Jeffrey C. Jackson, Christino Tamon, and Tomoyuki Yamakami. 'Quantum DNF learnability revisited'. In: *International Computing and Combinatorics Conference*. Springer. 2002, pp. 595–604. DOI: 10.1007/3-540-45655-4_63 (cit. on pp. 48, 50).

[141] Alp Atıcı and Rocco A. Servedio. 'Quantum Algorithms for Learning and Testing Juntas'. In: *Quantum Information Processing* 6.5 (2007), pp. 323–348. ISSN: 1570-0755. DOI: 10.1007/s11128-007-0061-6 (cit. on p. 48).

[142] Diego Ristè, Marcus P. Da Silva, Colm A. Ryan, Andrew W. Cross, Antonio D. Córcoles, John A. Smolin, Jay M. Gambetta, Jerry M. Chow, and Blake R. Johnson. 'Demonstration of quantum advantage in machine learning'. In: *npj Quantum Information* 3.1 (2017), pp. 1–5. DOI: 10.1038/s41534-017-0017-3 (cit. on pp. 48, 50, 51).

[143] Srinivasan Arunachalam, Alex B. Grilo, and Aarthi Sundaram. *Quantum hardness of learning shallow classical circuits*. Version 2. Sept. 19, 2019. arXiv: 1903.02840 [quant-ph] (cit. on p. 48).

[144] Nathan Linial, Yishay Mansour, and Noam Nisan. 'Constant depth circuits, Fourier transform, and learnability'. In: *Journal of the ACM (JACM)* 40.3 (1993), pp. 607–620. DOI: 10.1145/174130.174138 (cit. on pp. 48, 49).

[145] Yishay Mansour. 'Learning Boolean functions via the Fourier transform'. In: *Theoretical advances in neural computation and learning*. Springer, 1994, pp. 391–424. DOI: 10.1007/978-1-4615-2696-4_11 (cit. on p. 48).

[146] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014 (cit. on p. 48).

[147] Oded Goldreich and Leonid A Levin. 'A hard-core predicate for all one-way functions'. In: *Proceedings of the twenty-first annual ACM symposium on Theory of computing*. 1989, pp. 25–32. DOI: 10.1145/73007.73010 (cit. on p. 50).

[148] Oded Regev. 'On lattices, learning with errors, random linear codes, and cryptography'. In: *Journal of the ACM (JACM)* 56.6 (2009), pp. 1–40. DOI: 10.1145/1568318.1568324 (cit. on p. 51).

[149] Avrim Blum, Adam Kalai, and Hal Wasserman. 'Noise-tolerant learning, the parity problem, and the statistical query model'. In: *Journal of the ACM (JACM)* 50.4 (2003), pp. 506–519. DOI: 10.1145/792538.792543 (cit. on p. 51).

[150] Vadim Lyubashevsky. 'The Parity Problem in the Presence of Noise, Decoding Random Linear Codes, and the Subset Sum Problem'. In: *Approximation, randomization and combinatorial optimization. Algorithms and techniques*. Springer, 2005, pp. 378–389. DOI: 10.1007/11538462_32 (cit. on p. 51).

[151] Sanjeev Arora and Rong Ge. 'New Algorithms for Learning in Presence of Errors'. In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2011, pp. 403–415. DOI: 10.1007/978-3-642-22006-7_34 (cit. on p. 51).

[152] Kai-Min Chung and Han-Hsuan Lin. 'Sample Efficient Algorithms for Learning Quantum Channels in PAC Model and the Approximate State Discrimination Problem'. In: *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*. Ed. by Min-Hsiu Hsieh. Vol. 197. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 3:1–3:22. ISBN: 978-3-95977-198-6. DOI: 10.4230/LIPIcs.TQC.2021.3 (cit. on pp. 52, 53).

[153] Hsin-Yuan Huang, Richard Kueng, and John Preskill. 'Information-Theoretic Bounds on Quantum Advantage in Machine Learning'. In: *Physical Review Letters* 126.19 (2021), p. 190505. DOI: 10.1103/PhysRevLett.126.190505 (cit. on pp. 52, 53, 55, 56).

[154] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. *Quantum Algorithmic Measurement*. Version 2. July 21, 2021. arXiv: 2101.04634 [quant-ph] (cit. on pp. 52, 53).

[155] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. *Exponential separations between learning with and without quantum memory*. Version 1. Nov. 10, 2021. arXiv: 2111.05881 [quant-ph] (cit. on pp. 52, 53, 55, 56).

[156] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. *A Hierarchy for Replica Quantum Advantage*. Version 1. Nov. 10, 2021. arXiv: 2111.05874 [quant-ph] (cit. on pp. 52, 53).

[157] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean. *Quantum advantage in learning from experiments*. Version 1. Dec. 1, 2021. arXiv: 2112.00778 [quant-ph] (cit. on pp. 52, 53, 56).

[158] Jordan Cotler, Hsin-Yuan Huang, and Jarrod R. McClean. *Revisiting dequantization and quantum advantage in learning tasks*. Version 2. Dec. 6, 2021. arXiv: 2112.00811 [quant-ph] (cit. on p. 52).

[159] Philip D. Laird. *Learning from good and bad data*. Springer Science & Business Media, 2012 (cit. on p. 55).

[160] Klaus-Jochen Engel and Rainer Nagel. *One-Parameter Semigroups for Linear Evolution Equations*. Graduate Texts in Mathematics. Springer New York, 2006. ISBN: 9780387226422. DOI: 10.1007/b97696 (cit. on p. 59).

[161] Toby S. Cubitt, Jens Eisert, and Michael M. Wolf. 'The Complexity of Relating Quantum Channels to Master Equations'. In: *Communications in Mathematical Physics* 310.2 (2012), pp. 383–418. DOI: 10.1007/s00220-011-1402-y (cit. on p. 60).

[162] Michael M. Wolf, Jens Eisert, Toby S. Cubitt, and J. Ignacio Cirac. 'Assessing Non-Markovian Quantum Dynamics'. In: *Physical review letters* 101.15 (2008), p. 150402. DOI: 10.1103/PhysRevLett.101.150402 (cit. on p. 60).

[163] Emilio Onorati, Tamara Kohler, and Toby Cubitt. *Fitting quantum noise models to tomography data*. Version 2. July 9, 2021. arXiv: 2103.17243 [quant-ph] (cit. on p. 60).

[164] Ángel Rivas, Susana F. Huelga, and Martin B. Plenio. 'Entanglement and Non-Markovianity of Quantum Evolutions'. In: *Phys. Rev. Lett.* 105 (5 July 2010), p. 050403. DOI: 10.1103/PhysRevLett.105.050403 (cit. on p. 60).

[165] Heinz-Peter Breuer, Elsi-Mari Laine, and Jyrki Piilo. 'Measure for the Degree of Non-Markovian Behavior of Quantum Processes in Open Systems'. In: *Phys. Rev. Lett.* 103 (21 Nov. 2009), p. 210401. DOI: 10.1103/PhysRevLett.103.210401 (cit. on p. 60).

[166] Li Li, Michael J. W. Hall, and Howard M. Wiseman. 'Concepts of quantum non-Markovianity: A hierarchy'. In: *Physics Reports* 759 (2018), pp. 1–51. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2018.07.001 (cit. on p. 60).

[167] Ángel Rivas, Susana F. Huelga, and Martin B. Plenio. 'Quantum non-Markovianity: Characterization, quantification and detection'. In: *Reports on Progress in Physics* 77.9 (2014), p. 094001. ISSN: 0034-4885. DOI: 10.1088/0034-4885/77/9/094001 (cit. on p. 60).

[168] Heinz-Peter Breuer, Elsi-Mari Laine, Jyrki Piilo, and Bassano Vacchini. 'Colloquium: Non-Markovian dynamics in open quantum systems'. In: *Reviews of Modern Physics* 88.2 (2016), p. 021002. DOI: 10.1103/RevModPhys.88.021002 (cit. on p. 60).

[169] C.-F. Li, G.-C. Guo, and J. Piilo. 'Non-Markovian quantum dynamics: What does it mean?' In: *EPL (Europhysics Letters)* 127.5 (2019), p. 50001. ISSN: 0295-5075. DOI: 10.1209/0295-5075/127/50001 (cit. on p. 60).

Appendix A

Core Articles

A.1 Pseudo-dimension of quantum circuits

Pseudo-dimension of quantum circuits

Matthias C. Caro and Ishaun Datta

The pseudo-dimension is a complexity measure from classical learning theory. There, it is used to quantify the expressivity of real-valued function classes. In particular, bounds on the pseudo-dimension of a function class allow to derive generalization bounds for learning the function class. Such generalization bounds are central in statistical learning theory for understanding the training data requirements of machine learning models.

Quantum circuits are a central object of study in quantum computing. They are the natural quantum analogue of classical circuits. To ensure efficient implementability, one often considers quantum circuits that are 2-local, which means that every quantum gate acts on at most 2 qudits, and that have size (i.e., number of gates) and depth (i.e., number of gate layers) polynomial in the number of input qudits.

In this work, we propose the pseudo-dimension as a tool to quantify the complexity of 2-local quantum circuits. We establish pseudo-dimension bounds in terms of circuit depth and size. Moreover, we present two applications of our learning theory-inspired perspective, namely to the gate complexity of state preparation and to the learnability of quantum circuits.

After the introduction, in which we motivate the questions of the article, discuss related work, and summarize our results, Section 2 introduces basic notions of quantum information and statistical learning theory. In Section 3 of the paper, we present our setup and our main results. We first consider 2-local quantum circuits in which the layout of the 2-qudit gates is fixed, but the gates themselves can be arbitrary 2-qudit unitaries. We can view this as a quantum circuit with fixed architecture and variable gates. To such a circuit, we associate a class of $[0,1]$-valued functions by considering, for any fixed choice for the variable unitaries, the outcome probabilities of rank-1 projective measurements performed independently at the output of the circuit, upon input of the $|0\rangle$ state. In Theorem 2, we prove that the pseudo-dimension of this function class scales at worst as $\mathcal{O}(d^4 \cdot \gamma \log \gamma)$, with d the local dimension and γ the size of the circuit. That is, the pseudo-dimension grows at most polynomially in the qudit dimension and slightly superlinearly in the number of 2-local gates in the circuit. To show this result, we first establish a representation of functions in our class of interest in terms of a polynomial, whose rank depends on the number of gates in the circuit. We then combine this with a known bound on consistent sign assignments to a family of polynomials (Theorem 1 and Corollary 1) to obtain Theorem 2.

The remainder of Section 3 is concerned with extensions of Theorem 2 to different scenarios. First, we admit also quantum circuits with variable architecture (Theorem 3). Next, we allow for variable input states (Subsection 3.3). And third, we consider circuits consisting not of unitaries but of gates described by completely positive and trace-preserving maps (Theorem 4). Crucially, we prove that in all of these extensions, the pseudo-dimension of the respective function class is still upper bounded by a polynomial in the qudit dimension, the circuit depth, and the circuit size.

We continue in Section 4 with two applications of our pseudo-dimension bounds. On the one hand, we demonstrate a connection between complexity as measured by the pseudo-dimension

and the gate complexity of state preparation. More precisely, we construct a concrete set of pure n-qubit quantum states such that at least one of those states cannot be implemented by a 2-local quantum circuit with subexponential (in n) depth or size. As our set of candidate states has cardinality doubly exponential in n, in this case, our pseudo-dimension-based approach provides a constructive alternative to a more standard reasoning based on covering number arguments. On the other hand, we use classical generalization bounds in terms of the fat-shattering or the pseudo-dimension to show that 2-local quantum circuits of polynomial depth and size can be learned from polynomial-size training data, in which each example is a triple of input state, observed output measurement outcome, and corresponding measurement probability. This result constitutes a "pretty good" version of quantum process tomography.
I was significantly involved in finding the ideas and carrying out the scientific work of all parts of this article. I was in charge of writing the article, with the exception of the Appendix.

Permission to include:

Matthias C. Caro and Ishaun Datta.
Pseudo-dimension of quantum circuits.
Quantum Mach. Intell. 2, 14 (2020). https://doi.org/10.1007/s42484-020-00027-5.

Springer www.springer.com

Permissions

Get permission to reuse Springer Nature content

Springer Nature is partnered with the Copyright Clearance Center to meet our customers' licensing and permissions needs.

Copyright Clearance Center's RightsLink® service makes it faster and easier to secure permission for the reuse of Springer Nature content to be published, for example, in a journal/magazine, book/textbook, coursepack, thesis/dissertation, annual report, newspaper, training materials, presentation/slide kit, promotional material, etc.

Simply visit SpringerLink and locate the desired content;

Go to the article or chapter page you wish to reuse content from. (Note: permissions are granted on the article or chapter level, not on the book or journal level). Scroll to the botton of the page, or locate via the side bar, the "Reprints and Permissions" link at the end of the chapter or article.

Select the way you would like to reuse the content;

Complete the form with details on your intended reuse. Please be as complete and specific as possible ao as not to delay your permission request;

Create an account if you haven't already. A RightsLink account is different than a SpringerLink account, and is necessary to receive a licence regardless of the permission fee. You will receive your licence via the email attached to your RightsLink receipt;

Accept the terms and conditions and you're done!

For questions about using the RightsLink service, please contact Customer Support at Copyright Clearance Center via phone +1-855-239-3415 or +1-978-646-2777 or email springernaturesupport@copyright.com.

How to obtain permission to reuse Springer Nature content not available online on SpringerLink

Requests for permission to reuse content (e.g. figure or table, abstract, text excerpts) from Springer Nature publications currently not available online must be submitted in writing. Please be as detailed and specific as possible about what, where, how much, and why you wish to reuse the content.

Your contacts to obtain permission for the reuse of material from:

- books: bookpermissions@springernature.com
- journals: journalpermissions@springernature.com

Author reuse

Please check the Copyright Transfer Statement (CTS) or Licence to Publish (LTP) that you have signed with Springer Nature to find further information about the reuse of your content.

Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.
Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]'

If you are any doubt about whether your intended re-use is covered, please contact journalpermissions@springernature.com for confirmation.

Self-Archiving

- Journal authors retain the right to self-archive the final accepted version of their manuscript. Please see our self-archiving policy for full details:
https://www.springer.com/gp/open-access/authors-rights/self-archiving-policy/2124

- Book authors please refer to the information on this link:
https://www.springer.com/gp/open-access/publication-policies/self-archiving-policy

Quantum Machine Intelligence (2020) 2:14
https://doi.org/10.1007/s42484-020-00027-5

RESEARCH ARTICLE

Pseudo-dimension of quantum circuits

Matthias C. Caro[1] · **Ishaun Datta**[1,2]

Received: 25 April 2020 / Accepted: 7 October 2020 / Published online: 6 November 2020

Abstract

We characterize the expressive power of quantum circuits with the pseudo-dimension, a measure of complexity for probabilistic concept classes. We prove pseudo-dimension bounds on the output probability distributions of quantum circuits; the upper bounds are polynomial in circuit depth and number of gates. Using these bounds, we exhibit a class of circuit output states out of which at least one has exponential gate complexity of state preparation, and moreover demonstrate that quantum circuits of known polynomial size and depth are PAC-learnable.

Keywords Quantum computing · Computational learning theory · Complexity theory

1 Introduction

An important line of research in classical learning theory is characterizing the expressive power of function classes using complexity measures. Such complexity bounds can in turn be used to bound the size of training data required for learning. Among the most prominent of these are the Vapnik-Chervonenkis (VC) dimension introduced by Vapnik and Chervonenkis (1971). Other well-known measures are the pseudo-dimension due to Pollard (1984), the fat-shattering dimension due to Alon et al. (1997), the Rademacher complexities (see Bartlett and Mendelson 2002), and more generally covering numbers in metric spaces.

The goal of characterizing an object's expressive power also appears in different guises throughout quantum information. A well-known example is quantum state tomography. Aaronson (2007) related a variant of state tomography to a classical learning task whose fat-shattering dimension can be bounded using a particular function class related to the set of quantum states. Associated with this is a corresponding upper bound on sample complexity.

Aaronson (2007) observes that there is no analogous theorem for general quantum process tomography, but leaves as an open question whether there are restricted classes of operations that are information-efficiently learnable. We answer this question in the affirmative. In particular, we show that for quantum circuits with depth and size polynomial in the number of qubits, quantum process tomography is possible using only polynomially many examples.

Gate complexity of unitary implementation and state preparation are yet another example of how one may capture the richness of a function class that corresponds to a quantum computational process (see, e.g., Aaronson 2016). For unitary complexity, the challenge is to determine, e.g., how many two-qubit unitaries (i.e., two-qubit logical gates, in a computational setting) are required to implement a certain multi-qubit unitary (i.e., a quantum circuit). For the gate complexity of state preparation, it is to determine how many unitaries produce a certain multi-qubit state. An alternative perspective, adopted in this work, is to consider the expressive power of a set of circuits with a fixed number of unitaries.

In this work we describe a new way of applying complexity measures from classical learning, specifically pseudo-dimension, to quantum information. We associate with a quantum circuit a natural probabilistic function class describing the outcome probabilities of measurements performed on the circuit output. In this way, a function class corresponding to a quantum circuit can be studied with the classical tool of pseudo-dimension. Here, we show that the pseudo-dimension of such a class can be bounded in

✉ Matthias C. Caro
caro@ma.tum.de

Ishaun Datta
idatta@stanford.edu

[1] Department of Mathematics, Technical University of Munich, Munich, Germany

[2] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, USA

terms of a polynomial of the circuit depth and size. We also give two applications of these bounds, one for the gate complexity of quantum state preparation, the other in learnability of quantum circuits.

These findings are noteworthy not only because of the results themselves, but because we demonstrate the power of pseudo-dimension to gain insight into quantum computation. We hope that these tools may be applied to other problems in quantum computing in future work.

1.1 Related work

Aaronson (2007) showed that using the framework of PAC learning, one can introduce a variant of quantum state tomography and prove an upper bound on the required number of copies of the unknown state. This idea was developed further in Aaronson et al. (2018) and Aaronson (2018).

Motivated by Aaronson's work, Cheng et al. (2016) use pseudo-dimension and fat-shattering dimension to characterize the learnability of measurements, as a dual problem to learning the state. We apply this mathematical framework to study the problem of learning the circuit itself, in particular by offering a natural function class corresponding to a quantum circuit.

Rocchetto (2017) proved that stabilizer states, prevalent in error correction, are *computationally* efficiently learnable, establishing a connection between efficient classical simulability and computationally efficient learnability. This was realized experimentally for small optical systems in Rocchetto et al. (2019). Similarly, in Section 5 we pose as an open problem whether there are quantum operations that can be PAC-learned with modest computation, which could then in principle be demonstrated in an experiment.

In Chung and Lin (2018), the authors study the problem of PAC learning classes of functions with computational basis states as input and quantum output, possibly mixed. We highlight two main differences: first, whereas we assume the training data to be measurement statistics, Chung and Lin (2018) consider examples given as classical-quantum states. Thus, the two scenarios are not directly comparable. Our learning result yields a semi-classical strategy for the problem described in Chung and Lin (2018), though it is possibly suboptimal. Second, the learnability result of Chung and Lin (2018) is only for finite concept classes, whereas our result does not have this restriction. While Chung and Lin (2018) show learnability of quantum circuits with a finite gate set, we allow for arbitrary 2-qudit gates, i.e., a continuous gate set. Note that our corresponding notions of learnability differ.

While we take a formal approach to learning quantum circuits, others have studied learning unitaries numerically, e.g., with heuristics such as gradient descent (Kiani et al. 2020). Practical machine learning algorithms have also been used for state tomography by Torlai et al. (2018), and similar techniques could be applied to restricted classes of process tomography.

Another branch of quantum learning deals with whether quantum examples can decrease the information-theoretic complexity of learning a classical function. There are different flavors of this question, e.g., depending on whether learning is distribution-specific or distribution-independent. Arunachalam and de Wolf (2017) gives an overview of some of these aspects of quantum learning.

In classical learning theory, bounding the complexity measures of function classes (based on complexity-theoretic assumptions) has been studied widely. Goldberg and Jerrum (1995) derived an upper bound on the VC-dimension of a function class in terms of the runtime required by an algorithm implementing the elements of that class. Karpinski and Macintyre (1997) established an analogous bound for the function class implemented by a neural network (for various activation functions) in terms of the number of nodes and the number of programmable parameters of the network. Koiran (1996) demonstrated that by bounding the complexity of function classes implemented on a given architecture, one can lower bound the size of an architecture implementing a specific "hard" function.

1.2 Overview of results

We consider the general scenario in which one measures the output state of a 2-local qu*d*it quantum circuit, generating a probability distribution. We do not assume *geometric* locality, i.e., we do not assume that 2-qu*d*it unitaries act on neighboring qu*d*its. We show an upper bound on the pseudo-dimension of the distributions arising from these quantum circuits. By doing so, we provide insight into the complexity or "hardness" of the circuit and the output state that gives rise to the probability distribution. Below, we provide informal statements of the key results.

Theorem (Pseudo-dimension bounds, Informal) *Consider quantum circuits with fixed architecture, namely those for which the input qudits of the 2-qudit gates are specified, but the gates may vary subject to this constraint. That is, we allow for arbitrary 2-qudit unitaries, and in particular we do not restrict ourselves to a finite gate library.*

Parameterize a quantum circuit $\mathcal{N}$ *by its qudit dimension* d, *depth* δ, *and number of gates or size* γ.

Theorem 2: *For a suitable function class* $\mathcal{F}_{\mathcal{N}}$ *corresponding to the possible probability distributions formed by product measurements in the computational basis on the circuit output,* $Pdim(\mathcal{F}_{\mathcal{N}}) \leq \mathcal{O}(d^4 \cdot \gamma \log \gamma)$.

Consider quantum circuits with variable architecture, i.e., those for which the input qudits of the gates are not specified. For such circuits of depth δ and number of gates or size γ, one may similarly define function classes $\mathscr{F}_{\delta,\gamma}$ for circuits whose gates are unitaries, and $\mathscr{G}_{\delta,\gamma}$ for circuits whose gates are quantum operations, which describe the possible probability distributions formed by product measurements on the circuit output. Then,

Theorem 3: $Pdim(\mathscr{F}_{\delta,\gamma}) \leq \mathscr{O}(\delta \cdot d^4 \cdot \gamma^2 \log \gamma)$.

Theorem 4: $Pdim(\mathscr{G}_{\delta,\gamma}) \leq \mathscr{O}(\delta \cdot d^8 \cdot \gamma^2 \log \gamma)$.

All upper bounds are polynomial in the dimension d, the depth δ, and the size γ.

In Section 4.1, we demonstrate how to apply these complexity upper bounds to explicitly construct, for each $n \in \mathbb{N}$, a finite-but-large set of n-qubit quantum states, out of which at least one cannot be implemented by a 2-local qudit circuit of subexponential depth or size.

Theorem (Gate Complexity of State Preparation, Informal)

For any subset $C \subseteq \{|x0\rangle\}_{x\in\{0,1\}^n}$, define

$$|\psi_C\rangle = \begin{cases} \frac{1}{\sqrt{|C|}} \sum_{|x0\rangle\in C} |x0\rangle & \text{if } C \neq \emptyset \\ |0\rangle^{\otimes n} \otimes |1\rangle & \text{if } C = \emptyset. \end{cases}$$

If each state in $\{|\psi_C\rangle\}_C$ can be generated from the input state $|0\rangle^{\otimes(n+1)}$ by some circuit of depth δ and size γ, then $2^n \leq \mathscr{O}\left(\delta \cdot \gamma^2 \log \gamma\right)$. As a corollary, there exists at least one such C so that $|\psi_C\rangle$ requires a circuit exponential in depth and size.

Analogously to Aaronson (2007), in Section 4.2 we use our pseudo-dimension bounds to prove a relaxed variant of quantum process tomography, which following Aaronson's terminology can be called *pretty-good circuit tomography*:

Theorem (Learnability, Informal) *Given a circuit with depth Δ and size Γ, both polynomial in the number of qudits and known in advance to the learner, polynomially-many training examples, each a triple of input state, output measurement, and corresponding probability, suffice to learn the quantum operation implemented by a 2-local quantum circuit of depth Δ and size Γ.*

That is, for confidence δ, accuracy, ε, and error margins α and β, all in $(0, 1)$, a candidate circuit of depth Δ and size Γ that performs sufficiently well (in a sense made rigorous in Section 4.2*) on*

$$\mathscr{O}\left(\frac{1}{\varepsilon}\left(\Delta d^8 \Gamma^2 \log \Gamma \log^2\left(\frac{\Delta d^8 \Gamma^2 \log \Gamma}{(\beta-\alpha)\varepsilon}\right) + \log\frac{1}{\delta}\right)\right)$$

many samples will with probability at least $1 - \delta$ approximate the actual circuit from which the samples are drawn.

In this framework, each training example is a three-tuple of the input state, the observed measurement outcome, and the corresponding measurement probability. Alternately, one may take each training example as a two-tuple of the input state and the measurement outcome, whose probability is the corresponding measurement probability (see Aaronson 2007, Appendix 8).

We review the basics of quantum information, quantum computation, and classical learning theory in Section 2. We also discuss prior classical results as motivation. Section 3 contains our main results on the pseudo-dimension of quantum circuits and the respective proofs. In Section 4, we apply these results to fin lower bounds on the gate complexity of quantum state preparation and to a learning problem for quantum operations. We conclude with open questions in Section 5.

2 Preliminaries

As our readership includes both physicists and computer scientists, in this section we review the mathematical frameworks of quantum information theory and learning theory. Further details appear in the reference texts (Heinosaari and Ziman 2013; Nielsen and Chuang 2010).

2.1 Quantum information and computation

The most general descriptor of a d-level quantum system or statistical ensemble thereof is a density matrix, an element of

$$\mathscr{S}\left(\mathbb{C}^d\right) := \{\rho \in \mathbb{C}^{d\times d} \mid \rho \geq 0,\ \mathrm{tr}[\rho] = 1\}.$$

Here, $\rho \geq 0$ means that the matrix ρ is Hermitian and all its eigenvalues are non-negative. An important subset of density matrices is the set of pure states, which are one-dimensional projections. Following Dirac notation, we denote the projector onto the subspace spanned by a unit vector $|\psi\rangle \in \mathbb{C}^d$ by $|\psi\rangle\langle\psi|$. By the spectral theorem, every quantum state can be written as a convex combination of pure states, though this decomposition is not unique in general.

Central to the framework of quantum mechanics is the measurement, the mechanism by which one may observe properties of a quantum system. These are typically described by so-called positive-operator valued measures (POVMs). As we focus on measurements with a finite set of outcomes $\{i\}$, it suffices to think of measurements as collections of so-called effect operators $\{E_i\}_{i=1}^m$ with

$E_i \in \mathbb{C}^{d\times d}$, $0 \leq E_i \leq \mathbb{1}_d$, and $\sum_{i=1}^{m} E_i = \mathbb{1}_d$. We denote the set of effect operators by

$$\mathscr{E}\left(\mathbb{C}^d\right) := \{E \in \mathbb{C}^{d\times d} \mid 0 \leq E_i \leq \mathbb{1}_d\}.$$

Again, we highlight a special case: if we take an orthonormal basis $\{|\psi_i\rangle\}_{i=1}^d$ of $\mathbb{C}^d$, then the set $\{E_i = |\psi_i\rangle\langle\psi_i|\}_{i=1}^d$ is called a projective measurement.

Born's rule connects measurements to measurement outcomes: given a state characterized by a density operator, the effect operator has a corresponding probability $p_i = \mathrm{tr}[\rho E_i]$. Thus the requirement that the effect operators sum to the identity can be seen as probabilities summing to one. In the special case of pure state $\rho = |\psi\rangle\langle\psi|$ and projective measurement $\{E_i = |\psi_i\rangle\langle\psi_i|\}_{i=1}^d$, the probability of outcome i is $p_i = \mathrm{tr}[\rho E_i] = |\langle\psi|\psi_i\rangle|^2$.

So far we have described the components of static quantum theory. The dynamics of quantum states are described by so-called quantum operations, which we denote by

$$\mathscr{T}\left(\mathbb{C}^d\right) := \{ \quad T : \mathbb{C}^{d\times d} \to \mathbb{C}^{d\times d} \mid T \text{ is linear, completely positive, and trace-non-increasing}\}.$$

Here, a map T is completely positive if $T \otimes Id_n$ is positivity-preserving for every $n \in \mathbb{N}$. If $T \in \mathscr{T}\left(\mathbb{C}^d\right)$ is trace-preserving, we call T a quantum channel. An important example is the unitary quantum channel, $T(\rho) = U\rho U^*$ for some unitary $U \in \mathbb{C}^{d\times d}$.

Note that any element of $\mathscr{T}\left(\mathbb{C}^d\right)$ is a linear map between vector spaces of dimension d^2 and can thus be understood as a $d^2 \times d^2$ matrix.

2.2 Classical learning theory and complexity measures

Next we describe the "probably approximately correct" (PAC) model of learning, introduced and formalized by Vapnik and Chervonenkis (1971) and Valiant (1984). In (realizable) PAC learning for spaces X, Y and a concept class $\mathscr{F} \subseteq Y^X$, a learning algorithm receives as input labeled training data $\{(x_i, f(x_i))\}_{i=1}^m$ for some $f \in \mathscr{F}$, where the samples x_i are drawn independently according to some unknown probability distribution D on X that is unknown to the learner. Given the training examples, the goal of the learner is to approximate the unknown function f by a hypothesis function h, with high probability.

We can formalize this as follows: first, we introduce a loss function $\ell : Y \times Y \to \mathbb{R}_+$ to quantify the discrepancy between the hypothesis h and the function f. We call a concept class $\mathscr{F}$ PAC-learnable if there exists a learning algorithm $\mathscr{A}$ such that for every probability distribution D on X, $f \in \mathscr{F}$ and $\delta, \varepsilon \in (0, 1)$, running $\mathscr{A}$ on training data drawn according to D and f yields a hypothesis h such that $\mathbb{E}_{x\sim D}[\ell(h(x), f(x))] \leq \varepsilon$ with probability $\geq 1 - \delta$ (with regard to the choice of training data). Moreover, we quantify the minimum amount of training data that an algorithm $\mathscr{A}$ needs to meet the above conditions by a map $m_{\mathscr{F}} : (0, 1) \times (0, 1) \to \mathbb{N}$, $(\delta, \varepsilon) \mapsto m(\delta, \varepsilon)$, the so-called sample complexity of $\mathscr{F}$. We focus on proper learning, in which the learning algorithm must output as its hypothesis an element of the concept class, i.e., we require $h \in \mathscr{F}$.

A standard approach to assessing learnability is to characterize the complexity of the respective concept class $\mathscr{F}$. Many such complexity measures are used, the most common being the VC-dimension for binary-valued function classes $\mathscr{F} \subseteq \{0, 1\}^X$, named after its progenitors (Vapnik and Chervonenkis 1971). This combinatorial parameter can be shown to fully characterize the learnability: a concept class $\mathscr{F} \subseteq \{0, 1\}^X$ is PAC-learnable (w.r.t. the 0-1-loss) if and only if the VC-dimension of $\mathscr{F}$ is finite. Moreover, the sample complexity of PAC learning $\mathscr{F}$ can be expressed in terms of its VC-dimension (see Blumer et al. 1989; Hanneke 2016).

In this work, we employ a widely used extension of the VC-dimension to real-valued concept classes:

Definition 1 *(Pseudo-dimension* (Pollard 1984)*)* Let $\mathscr{F} \subseteq \mathbb{R}^X$ be a real-valued concept class. A set $\{x_1, ..., x_k\} \subseteq X$ is pseudo-shattered by $\mathscr{F}$ if there are $y_1, ..., y_k \in \mathbb{R}$ such that for any $C \subseteq \{1, ..., k\}$ there is an $f_C \in \mathscr{F}$ such that for all $1 \leq i \leq k$, $i \in C$ if and only if $f_C(x_i) \geq y_i$.

The pseudo-dimension of $\mathscr{F}$ is defined to be

$$\mathrm{Pdim}(\mathscr{F}) := \sup\{n \in \mathbb{N}_0 \mid \exists S \subseteq X \text{ s.t. } |S| = n \text{ and } S \text{ is pseudo-shattered by } \mathscr{F}\}.$$

Alternatively, one can express the pseudo-dimension in terms of the VC-dimension. Namely,

$$\mathrm{Pdim}(\mathscr{F}) = \mathrm{VC}(\{X \times \mathbb{R} \ni (x, y) \mapsto \mathrm{sgn}(f(x) - y) \mid f \in \mathscr{F}\}).$$

Here, the VC-dimension for a function class $\mathscr{H} \subseteq \{\pm 1\}^Z$ is defined as

$$\mathrm{VC}(\mathscr{H}) := \sup\{n \in \mathbb{N}_0 \mid \exists z_1, \ldots, z_n \in Z \text{ s.t. } \forall b \in \{\pm 1\}^n \; \exists h_b \in \mathscr{H} \text{ s.t. } \forall i : h_b(z_i) = b_i\}.$$

There is also a scale-sensitive version of the pseudo-dimension:

Definition 2 *(Fat-Shattering Dimension* (Alon et al. 1997)*)* Let $\mathscr{F}$ be a real-valued concept class and let $\alpha > 0$. A set $\{x_1, ..., x_k\} \subseteq X$ is α-fat-shattered by $\mathscr{F}$ if there are $y_1, ..., y_k \in \mathbb{R}$ such that for any $C \subseteq \{1, ..., k\}$ there is an $f_C \in \mathscr{F}$ such that for all $1 \leq i \leq k$:

1. $i \notin C \Rightarrow f_C(x_i) \leq y_i - \alpha$ and

2. $i \in C \Rightarrow f_C(x_i) \geq y_i + \alpha$.

The α-fat-shattering dimension of $\mathscr{F}$ is defined to be

$$fat_{\mathscr{F}}(\alpha) := \sup\{n \in \mathbb{N}_0 | \quad \exists S \subseteq X \text{ s.t. } |S| = n \wedge \text{S is } \alpha\text{-fat-shattered by } \mathscr{F}\}.$$

Note that, trivially, $fat_{\mathscr{F}}(\alpha) \leq \text{Pdim}(\mathscr{F})$ holds for every $\alpha > 0$ and for every real-valued function class $\mathscr{F}$.

Sample complexity upper bounds for $[0, 1]$-valued function classes in terms of the fat-shattering dimension have been proved in Bartlett and Long (1998) and Anthony and Bartlett (2000).

3 Pseudo-dimension bounds for quantum circuits

We now formulate how to characterize the expressive power of quantum circuits. In particular, we consider circuits with n input registers of qu$d$its, size (i.e., number of gates) γ, and depth (i.e., number of layers) δ. More precisely, we consider circuits composed of two-qudit unitaries, i.e., logical gates with two inputs. Note that two-qudit gates include one-qudit gates. We assume that gates in the same layer and acting on disjoint pairs of qudits can act in parallel. Additionally, we assume that each qu$d$it is acted upon by at least one gate, else it effectively does not participate in the circuit.

In this section, we assign function classes to quantum circuits and then derive bounds on the pseudo-dimension of these function classes, in terms of the number of qudits and the size and depth of the circuits. First, we fix quantum circuit structure and inputs, varying only the entries of the unitary gates and thereby the resulting function. Then, we broaden our scope to variable circuit architectures, variable inputs, and circuits whose "gates" are general quantum operations.

An important tool that will recur throughout our work is the following result on polynomial sign assignments, used in Goldberg and Jerrum (1995) to derive VC-dimension bounds from computational complexity.

Theorem 1 (Warren 1968, Theorem 3) *Let $\{p_1, \ldots, p_m\}$ be a set of real polynomials in n variables with $m \geq n$, each of degree at most $d \geq 1$. Then the number of consistent non-zero sign assignments to $\{p_1, \ldots, p_m\}$ is at most $\left(\frac{4edm}{n}\right)^n$.*

Here, e is Euler's number and a "consistent non-zero sign assignment" to a set of polynomials $\{p_1, \ldots, p_m\}$ is a vector $b \in \{\pm 1\}^m$ s.t. there exist $x_1, \ldots, x_n \in \mathbb{R}$ for which it holds that $\text{sgn}(p_i(x_1, \ldots, x_n)) = b_i$ for all $1 \leq i \leq m$.

The following implication of Theorem 1 for consistent but not necessarily non-zero sign assignments (which we define as above, but with $b \in \{-1, 0, 1\}^m$) to sets of polynomials was observed in Goldberg and Jerrum (1995, Corollary 2.1).

Corollary 1 *Let $\{p_1, \ldots, p_m\}$ be a set of real polynomials in n variables with $m \geq n$, each of degree at most $d \geq 1$. Then the number of consistent sign assignments to $\{p_1, \ldots, p_m\}$ is at most $\left(\frac{8edm}{n}\right)^n$.*

Proof (Sketch) This can be obtained by applying Theorem 1 to the set $\{p_1 + \varepsilon, p_1 - \varepsilon, \ldots, p_m + \varepsilon, p_m - \varepsilon\}$ with $\varepsilon > 0$ chosen sufficiently small. □

3.1 Fixed circuit structure

Suppose we fix the architecture of a quantum circuit of depth δ and size γ. Specifically, we restrict our attention to 2-local quantum circuits, i.e., circuits whose logical gates have support on two qudits, not necessarily neighboring each other (see Fig. 1). "Fixed architecture" means that we specify the positions of the two-qudit unitaries, namely their order and which qudits they act on. Though the unitaries' positions are fixed, we may vary the entries of the unitaries themselves. Here, we allow for arbitrary 2-qu$d$it unitaries. In particular, we do not restrict ourselves to a finite gate library. Can we bound the pseudo-dimension of the function class of measurement probability distributions that this circuit generates? And how does the bound depend on d (the dimensionality of the qudits), δ and γ?

To formalize this question: let $n \in \mathbb{N}$ be the number of qudits, $d \in \mathbb{N}$ be their dimensionality, and $\mathscr{N}$ be a fixed quantum circuit architecture of depth δ and size γ acting on n qudits. We enumerate the positions of the two-qudit unitaries in $\mathscr{N}$ by tuples (i, j) with $1 \leq i \leq \delta$ denoting

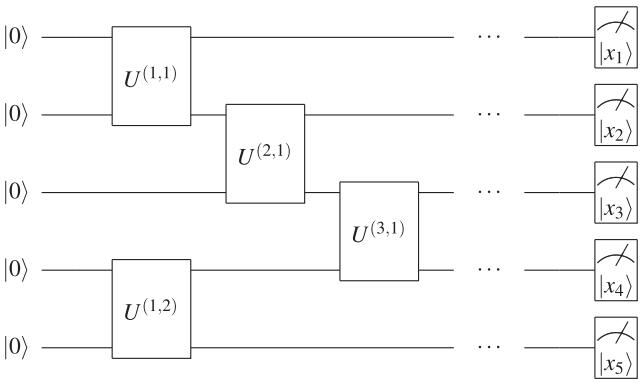


Fig. 1 An example 2-local circuit. $U^{(i,j)}$ denotes the jth 2-qudit unitary in the ith layer of the circuit

the layer and $1 \leq j \leq \gamma_i$ the position of the unitary among all the unitaries inside layer i, where w.l.o.g. we count from top to bottom and take into account only the first qudit on which a unitary acts.

Note that $\sum_{i=1}^{\delta} \gamma_i = \gamma$, and trivially $\gamma_i \leq \gamma$ and $\gamma_i \leq \frac{n}{2}$, as we assume that every qudit is acted upon by at least one gate. We write the unitary at position (i, j) as $U^{(i,j)}$. These constitute the "free parameters" which we can vary in order to make the quantum circuit perform different tasks. The overall unitary implemented by $\mathscr{N}$ when plugging in the unitaries $\{U^{(i,j)}\}_{1\leq i\leq\delta, 1\leq j\leq\gamma_i}$ at the respective positions we denote by $U_{\mathscr{N}|\{U^{(i,j)}\}}$. Note that $U_{\mathscr{N}|\{U^{(i,j)}\}}$ strongly depends on the two-qudit unitaries that are plugged into the architecture, but sometimes we will suppress this dependence and simply write $U_{\mathscr{N}}$ for notational ease.

The quantum circuit $\mathscr{N}$ now gives rise to the following set of output states:

$$\mathscr{S}_{\mathscr{N}}\left((\mathbb{C}^d)^{\otimes n}\right) := \left\{U_{\mathscr{N}|\{U^{(i,j)}\}} |0\rangle^{\otimes n} \middle|\ U^{(i,j)} \in \mathscr{U}\left((\mathbb{C}^d)^{\otimes 2}\right)\right\}.$$

These output states in turn give rise to a function class of measurement probability distributions with regard to product measurements:

$$\mathscr{F}_{\mathscr{N}} := \left\{f : X \to [0,1] \mid \exists |\psi\rangle \in \mathscr{S}_{\mathscr{N}}\left((\mathbb{C}^d)^{\otimes n}\right) : f(x) = |\langle x|\psi\rangle|^2\right\},$$

where we take $X = S_d \times \ldots \times S_d$ to be the Cartesian product of n unit spheres of $\mathbb{C}^d$.

The main insight of this subsection is the following:

Theorem 2 *With the notation and assumptions from above, it holds that* $Pdim(\mathscr{F}_{\mathscr{N}}) \leq 8d^4 \cdot \gamma \cdot \log(16e \cdot \gamma)$.

Here and throughout the paper, log denotes the logarithm to base 2.

To prove this result, we provide the following.

Lemma 1 *With the notation and assumptions from above, there exists a polynomial* $p_{\mathscr{N}}$ *with real coefficients, in* $2\gamma d^4 + 2dn$ *real variables of degree* $\leq 2(\gamma + n)$ *such that every* $f \in \mathscr{F}_{\mathscr{N}}$ *can be obtained from* $p_{\mathscr{N}}$ *by fixing values for the first* $2\gamma d^4$ *variables. Moreover, in each term of p, the degree in the first* $2\gamma d^4$ *real variables is* $\leq 2\gamma$ *and the degree in the last* $2dn$ *real variables is* $\leq 2n$.

Notably, there is no explicit dependence on depth δ.

Proof We first observe that

$$|\langle x|U_{\mathscr{N}}|0\rangle^{\otimes n}|^2 = |\langle 0|^{\otimes n} U_{\mathscr{N}}^{\dagger}|x\rangle|^2.$$

We study this expression in a layer-wise analysis. When reading the circuit from right to left, the state that enters layer δ is transformed by the unitary $\bigotimes_{j=1}^{\gamma_\delta} U^{(\delta,j)\dagger}$ such that each amplitude of the state after the δth layer is a linear combination of the amplitudes of $|x\rangle$, where each coefficient is a multilinear monomial of degree γ_δ in some of the $\gamma_\delta \cdot d^4$ complex entries of the $\{U^{(\delta,j)\dagger}\}_{1\leq j\leq\gamma_\delta}$.

By iterating this reasoning, we see that the state after the $(\delta - i)$th layer has amplitudes which are given by a linear combination of the amplitudes of $|x\rangle$, where each coefficient is a multilinear polynomial of degree $\leq \sum_{k=0}^{i} \gamma_{\delta-k}$ in (some of) the entries of the unitaries $\{U^{(\delta-k,j_k)\dagger}\}_{0\leq k\leq i, 1\leq j_k\leq\gamma_k}$.

In particular, the $|0\rangle^{\otimes n}$-amplitude of the state $U_{\mathscr{N}}^{\dagger}|x\rangle$ can be written as a linear combination of the amplitudes of $|x\rangle$, where each coefficient is given by a multilinear polynomial $q_{\mathscr{N}}$ of degree $\leq \sum_{k=0}^{\delta} \gamma_{\delta-k} = \gamma$ in (some of) the $\gamma \cdot d^4$ complex entries of the unitaries $\{U^{(i,j_i)\dagger}\}_{0\leq i\leq\delta, 1\leq j_i\leq\gamma_i}$.

Recalling that the probability of observing outcome $|0\rangle^{\otimes n}$ is the square of the absolute value of the corresponding amplitude of $|x\rangle$, we obtain from the polynomial $q_{\mathscr{N}}$ a polynomial $p_{\mathscr{N}} = |q_{\mathscr{N}}|^2$ that describes the output probabilities. As $q_{\mathscr{N}}$ has degree at most γ in the $\gamma \cdot d^4$ complex parameters of the unitaries, $p_{\mathscr{N}}$ has degree at most 2γ in the corresponding $2\gamma \cdot d^4$ real parameters. Fixing these $2\gamma d^4$ parameters corresponds to fixing the circuit, and therefore one may obtain every $f \in \mathscr{F}_{\mathscr{N}}$ by fixing these parameters in $p_{\mathscr{N}}$.

Moreover, $p_{\mathscr{N}}$ is a polynomial in the $2dn$ real parameters which give rise to the amplitudes of $|x\rangle$. (Here, the assumption that $|x\rangle$ is a product state enters.) As each such amplitude has degree $\leq n$ in the $2dn$ complex parameters, the degree of $p_{\mathscr{N}}$ in these real parameters is at most $2n$. □

Remark 1 We formulate the result only for measurement operators consisting of tensor products of 1-dimensional projections, and continue to do so throughout this manuscript. For $x \in X$, we can write $|x\rangle = \bigotimes_{i=1}^{n} \left(\sum_{j=0}^{d-1} \alpha_j^{(i)} |j\rangle\right)$, so we associate dn complex variables with x. That each amplitude of $|x\rangle$ can be written as a product of n complex parameters gives rise to the upper bound of n in the degree.

We could instead look at more general measurement operators consisting of 1-dimensional projections without requiring product structure, i.e., entangled measurements. In this scenario, we would write $|x\rangle = \sum_{z\in\{0,\ldots,d-1\}^n} x_z |z\rangle$, associating d^n complex variables with x. In this setup, each amplitude of x is simply a polynomial of degree 1 in these complex variables.

As we fix the variables corresponding to x and y in the shattering assumption that appears in our proof of Theorem 2, their corresponding degrees are not relevant to

our argument; only the degree in the entries of the unitaries enters our analysis. Therefore, both product measurements or entangled measurements lead to the same pseudo-dimension bound. This is due to the fact that allowing for entangled measurements changes the set of allowed inputs but not the function class itself.

Now that we have established Lemma 1, we can prove Theorem 2 with reasoning analogous to that in Goldberg and Jerrum (1995).

Proof (Theorem 2) Let $\{(x_i, y_i)\}_{i=1}^m \subseteq X \times \mathbb{R}$ be such that for every $C \subseteq \{1, \ldots, m\}$ there exists $f_C \in \mathscr{F}_{\mathscr{N}}$ such that $f_C(x_i) - y_i \geq 0$ if and only if $i \in C$.

By Lemma 1, there exists a polynomial $p_{\mathscr{N}}$ in $2\gamma d^4 + 2dn$ real variables of degree $\leq 2(\gamma + n)$ such that for every $C \subseteq \{1, \ldots, m\}$ there exists an assignment Ξ_C to the first $2\gamma d^4$ variables of $p_{\mathscr{N}}$ such that $p_{\mathscr{N}}(\Xi_C, x_i) - y_i \geq 0$ if and only if $i \in C$.

In particular, this implies (using the "moreover" part of Lemma 1) that the set $\mathscr{P} = \{p_{\mathscr{N}}(\cdot, x_i) - y_i\}_{i=1}^m$ is a set of m polynomials of degree $\leq 2\gamma$ in $2\gamma d^4$ real variables that has at least 2^m different consistent sign assignments.

We now claim that $m \leq 8d^4 \cdot \gamma \cdot \log(16e \cdot \gamma)$. If $m < 2\gamma d^4$, this holds trivially. Hence, w.l.o.g. $m \geq 2\gamma d^4$. So by Corollary 1, we have

$$2^m \leq \left(\frac{8e \cdot 2\gamma \cdot m}{2\gamma d^4}\right)^{2\gamma d^4}.$$

Taking logarithms now gives

$$m \leq 2\gamma d^4 \left(\log(16e \cdot \gamma) + \log\left(\frac{m}{2\gamma d^4}\right)\right).$$

Now we distinguish cases. If $16e \cdot \gamma \geq \frac{m}{2\gamma d^4}$, then the above immediately implies $m \leq 4\gamma d^4 \cdot \log(16e \cdot \gamma)$. If $16e \cdot \gamma \leq \frac{m}{2\gamma d^4}$, then we obtain $m \leq 4\gamma d^4 \cdot \log\left(\frac{m}{2\gamma d^4}\right)$, which in turn implies $m \leq 8\gamma d^4$. In both cases we have $m \leq 8d^4 \cdot \gamma \cdot \log(16e\gamma)$. By definition of the pseudo-dimension, we conclude $\mathrm{Pdim}(\mathscr{F}_{\mathscr{N}}) \leq 8d^4 \cdot \gamma \cdot \log(16e\gamma)$, as claimed. □

The attentive reader may notice that we do not explicitly refer to the unitarity assumption in our reasoning; our argument mainly uses linearity. This already hints at a generalization to quantum circuits not of unitaries but of operations, which we will describe in Section 3.4. In that subsection, we will also see how the unitarity assumption implicit in this proof produces a better upper bound than in the general setting of quantum operations.

Remark 2 We formulate our bounds in terms of the pseudo-dimension, not its scale-sensitive version called fat-shattering dimension, even though the latter is more commonly used in classical learning. In our scenario, however, the pseudo-dimension and the fat-shattering dimension effectively coincide. This is because we could apply our reasoning for general matrices instead of only unitaries in the setting of Theorem 2 as well and achieve the same bounds. In that case, however, the resulting real-valued function class is closed under scalar multiplication with non-negative scalars and it follows from the definition that for such classes, the fat-shattering dimension equals the pseudo-dimension.

3.2 Variable circuit structure

Whereas in the previous subsection we fixed a quantum circuit architecture and only varied the entries of the two-qudit unitaries plugged into this structure, we now additionally vary the structure of the quantum circuit architecture itself and consider the complexity of the class of all quantum circuits of a given depth and size. Once again, we consider 2-local quantum circuits, i.e., circuits with one- and two-qudit gates acting on arbitrary pairs of qudits.

The class of states which is of relevance in this analysis is

$$\mathscr{S}_{\delta,\gamma}\left((\mathbb{C}^d)^{\otimes n}\right) := \{|\psi\rangle \mid \exists \text{ quantum circuit } \mathscr{N} \text{ of depth } \delta \text{ and size } \gamma \text{ such that } |\psi\rangle \in \mathscr{S}_{\mathscr{N}}\left((\mathbb{C}^d)^{\otimes n}\right)\}.$$

Again, this set of states gives rise to a function class via

$$\mathscr{F}_{\delta,\gamma} := \{f : X \to [0,1] \mid \exists |\psi\rangle \in \mathscr{S}_{\delta,\gamma}\left((\mathbb{C}^d)^{\otimes n}\right) : f(x) = |\langle x|\psi\rangle|^2\},$$

where X is as above given by $X = S_d \times \ldots \times S_d$. As before, we want to bound the pseudo-dimension of this function class.

We summarize the result of this subsection in the following:

Theorem 3 *With the notation and assumptions from above, it holds that* $Pdim(\mathscr{F}_{\delta,\gamma}) \leq \mathscr{O}(\delta \cdot d^4 \cdot \gamma^2 \log \gamma)$.

As with Theorem 2, the main step towards this result consists of relating the functions appearing in $\mathscr{F}_{\delta,\gamma}$ to polynomials. The difference here is that we must upper bound the number of polynomials, as below.

Lemma 2 *With the notation and assumptions from above, there exists a set* $\mathscr{P}_{\delta,\gamma}$ *of polynomials with real coefficients, in* $2\gamma d^4 + 2dn$ *real variables of degree* $\leq 2(\gamma + n)$ *such that for every* $f \in \mathscr{F}_{\delta,\gamma}$ *there exists a polynomial* $p \in \mathscr{P}_{\delta,\gamma}$

such that f can be obtained from p by fixing values for the first $2\gamma d^4$ variables, and such that

$$|\mathscr{P}_{\delta,\gamma}| \leq \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^{\delta}.$$

Moreover, in each term of $p \in \mathscr{P}_{\delta,\gamma}$ the degree in the first $2\gamma d^4$ real variables is $\leq 2\gamma$ and the degree in the last $2dn$ real variables is $\leq 2n$.

Proof There are at most $\frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}$ ways to assign them among the δ layers. The term $\frac{\gamma!}{(\gamma-\delta)!}$ counts assigning a single gate to each layer, to ensure that there are no trivial (empty) layers. Having assigned each layer one gate, the remaining $\gamma-\delta$ gates may be distributed to any of the δ layers.

Next, we bound the number of ways of assigning qudits to the circuit layers, so that the qudits are inputs to the fixed-position unitaries. For our purposes, it suffices to crudely upper bound this by $n!$ for each single layer and thus by $(n!)^{\delta}$ overall. Hence, there are at most

$$\frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^{\delta}$$

different quantum circuit architectures. The proof is completed by applying Lemma 1 to every such quantum circuit architecture. □

Now that we have established Lemma 2, we can prove Theorem 3 by reasoning analogous to that in Goldberg and Jerrum (1995) (see the Appendix for the proof of Theorem 3).

3.3 Extension to circuits with variable inputs

We now modify the results of Sections 3.1 and 3.2 to allow not only for the fixed input $|0\rangle^{\otimes n}$, but also for variable input. This is of use, for instance, in Section 4.2, in which we consider the PAC-learnability of quantum circuits (of unitary gates or more general quantum channels). In that context, allowing variable input amounts to learning the entire quantum circuit, rather than just its action on $|0\rangle^{\otimes n}$. This is necessary in order to meaningfully compare the learning problem in Section 4.2 to exact circuit tomography.

To consider variable input states, we define the following function classes, analogously to those in Sections 3.1 and 3.2:

$$\mathscr{F}'_{\mathscr{N}} := \{ \quad f : X \times Y \to [0,1] \mid \exists U_{\mathscr{N}|\{U^{(i,j)}\}},$$
$$U^{(i,j)} \in \mathscr{U}\left((\mathbb{C}^d)^{\otimes 2}\right) : f(x,y) = |\langle x|U_{\mathscr{N}}|y\rangle|^2\},$$

where Y can be taken as the computational basis states $\{0,1,...,d-1\}^n$, or more generally as $Y = X = S_d \times ... \times S_d$.

Lemma 3 *With the notation and assumptions from above the following holds: There exists a polynomial $p'_{\mathscr{N}}$ in $2\gamma d^4 + 4dn$ real variables of degree $\leq 2\gamma + 4n$ such that every $f \in \mathscr{F}'_{\mathscr{N}}$ can be obtained from $p'_{\mathscr{N}}$ by fixing values for the first $2\gamma d^4$ variables. Moreover, in each term of $p'_{\mathscr{N}}$ the degree in the first $2\gamma d^4$ real variables is $\leq 2\gamma$, the degree in the $2dn$ real variables corresponding to $x \in X$ is $\leq 2n$, and the degree in the $2dn$ real variables corresponding to $y \in Y$ is $\leq 2n$.*

Proof Consider the product state input $|y\rangle = \sum_z y_z|z\rangle$. As we consider product states, each y_z is a product of n complex parameters. Following the same reasoning as before, for a fixed $z \in \{0,\ldots,d-1\}$, $\langle z|U_{\mathscr{N}}|x\rangle$ is a multilinear polynomial $q^z_{\mathscr{N}}$. Then, the amplitude $\langle y|U_{\mathscr{N}}|x\rangle$ is

$$q'_{\mathscr{N}}(x,y) = \langle y|U_{\mathscr{N}}|x\rangle = \sum_{z\in\{0,1,\ldots,d-1\}^n} \overline{y_z}\,\langle z|U_{\mathscr{N}}|x\rangle$$
$$= \sum_{z\in\{0,1,\ldots,d-1\}^n} \overline{y_z}\, q^z_{\mathscr{N}}(x).$$

In the above equation, $q'_{\mathscr{N}}(x,y)$ has degree at most n in y, and so upon squaring the amplitude $q'_{\mathscr{N}}(x,y)$ to obtain $p'_{\mathscr{N}}(x,y)$ as in Lemma 1, we have a degree at most $2n$ in the $2dn$ real variables corresponding to y. The rest follows from Lemma 1. □

The bound from Theorem 2 still holds for the case of variable circuit input, with the proof proceeding almost identically upon replacing Lemma 1 by Lemma 3. The $2d \cdot n$ additional variables that arise from the polynomial y-dependence do not alter the bound because we fix the values of these variables in the pseudo-shattering assumption.

3.4 Extension to circuits of quantum operations

We finish this section by describing an extension of Theorems 2 and 3 to the case of circuits of quantum operations, instead of only unitaries. This generalization is relatively straightforward because the decisive property of unitaries used in our previous proofs was not the preservation of inner products, but rather linearity. This setting is useful to, e.g., describe circuits with imperfect gates. Rather than consider a logical gate that implements a unitary exactly, each gate can instead be considered a quantum operation that executes the desired unitary with some probability, and, e.g., depolarizes input qudits with some probability. (Other noise models are of course possible.) Note that although quantum operations can, by Stinespring's dilation theorem, be viewed as subsystem dynamics of a larger, unitarily evolving system, if we only

have access to measurement data for the subsystem then we cannot directly apply our result for the unitary case.

We use analogous notation to that introduced at the beginning of Section 3.1, writing $T_{\mathscr{N}|\{T^{(i,j)}\}}$ for the overall quantum operation implemented by $\mathscr{N}$ when plugging the two-qudit quantum operations $\{T^{(i,j)}\}_{1\leq i\leq\delta,1\leq j\leq\gamma_i}$ into the respective positions of the quantum circuit.

The quantum circuit $\mathscr{N}$ (of operations) now gives rise to the set of output states

$$\mathscr{D}_{\mathscr{N}}\left((\mathbb{C}^d)^{\otimes n}\right) := \{T_{\mathscr{N}|\{T^{(i,j)}\}}(|0^n\rangle\langle 0^n|) \mid T^{(i,j)} \in \mathscr{T}\left((\mathbb{C}^d)^{\otimes 2}\right)\},$$

where we write $|0^n\rangle = |0\rangle^{\otimes n}$, so $|0^n\rangle\langle 0^n| = (|0\rangle\langle 0|)^{\otimes n}$.

By taking into account all possible quantum circuits of size γ and depth δ, we obtain

$$\mathscr{D}_{\delta,\gamma}\left((\mathbb{C}^d)^{\otimes n}\right) := \{\rho \mid \exists \text{ circuit } \mathscr{N} \text{ of two-qudit operations of size } \gamma \text{ and depth } \delta \text{ such that } \rho \in \mathscr{D}_{\mathscr{N}}\left((\mathbb{C}^d)^{\otimes n}\right)\}.$$

These states now yield again a p-concept class

$$\mathscr{G}_{\delta,\gamma} := \{f : X \to [0,1] \mid \exists \rho \in \mathscr{D}_{\delta,\gamma}\left((\mathbb{C}^d)^{\otimes n}\right) : f(x) = \langle x|\rho|x\rangle\}.$$

In this scenario, we show:

Theorem 4 *With the notation and assumptions from above, it holds that* $Pdim(\mathscr{G}_{\delta,\gamma}) \leq \mathscr{O}(\delta \cdot d^8 \cdot \gamma^2 \log \gamma)$.

Proof We only sketch the reasoning, as it is similar to that in the proof of Theorem 3. We first need to establish an analogue of Lemma 2. To this end, observe that a quantum operation acting on two-qudit states can be interpreted as a $d^4 \times d^4$ matrix with complex entries. Moreover, we may write

$$\begin{aligned}\langle x|T_{\mathscr{N}}(|0^n\rangle\langle 0^n|)|x\rangle &= \mathrm{tr}[T_{\mathscr{N}}(|0^n\rangle\langle 0^n|)|x\rangle\langle x|] \\ &= \mathrm{tr}[|0^n\rangle\langle 0^n|T^*_{\mathscr{N}}(|x\rangle\langle x|)] \\ &= \langle 0^n|T^*_{\mathscr{N}}(|x\rangle\langle x|)|0^n\rangle,\end{aligned}$$

where $T^*_{\mathscr{N}}$ denotes the adjoint operation of $T_{\mathscr{N}}$ with regard to the Hilbert-Schmidt inner product.

As before, we can do a layer-wise analysis of the transformation of $|x\rangle\langle x|$ and observe that the entries of the (sub-normalized) density matrix after a layer can be written as linear combinations of the entries of the (sub-normalized) density matrix before the layer. Moreover, the coefficients can be written as multilinear polynomials with the degree determined by the number of two-qudit operations in the layer. Hence, we obtain the result of Lemma 1 with d^8 instead of d^4. The bound on the number of different quantum circuit architectures can be derived in exactly the same way as before, so the analogue of Lemma 2 holds, completing the proof of the theorem. □

Theorem 4 and its proof sketch also help to elucidate the relevance of the unitarity assumption in Theorems 2 and 3. Unitarity justifies our restriction to pure states, but in other respects Theorems 2 and 3 do not exploit unitary. The difference between Theorems 3 and 4 amounts to the size of the matrices that represent the unitaries or quantum operations.

4 Applications

In this section, we explore two different applications of our pseudo-dimension upper bounds. First, we employ the pseudo-dimension to exhibit a large but finite discrete set of quantum states, out of which at least one is hard to implement in the sense that preparing it requires exponentially many 2-qubit unitaries. Second, we combine the pseudo-dimension bound with results from the theory of p-concept learning to derive the PAC-learnability of quantum circuits.

4.1 Lower bounds on the gate complexity of quantum state preparation

It is well known that almost all n-qubit unitaries require an exponential (in n) number of 2-qubit unitaries to be implemented. Similarly, almost all pure n-qubit states require an application of exponentially (in n) many 2-qubit unitaries to be generated from the $|0\rangle^{\otimes n}$ state (see, e.g., Nielsen and Chuang 2010). However, in neither case are there explicit examples of unitaries or states saturating this exponentiality bound (see Aaronson 2016 for more information on the gate complexity of unitary implementation and state preparation). We will use the pseudo-dimension as a tool to exhibit a discrete set of pure qubit states such that at least one of them requires exponentially many 2-qubit unitaries to be generated from $|0\rangle^{\otimes n}$.

The drawback of our result is that the size of this set is 2^{2^n} and thus unsatisfyingly large. By relatively simple deliberations this size can be reduced by an order of 2^n elements, though this is negligible compared to the overall size.

We now describe the construction of the candidate set of states. For a subset $C \subseteq \{|x0\rangle\}_{x\in\{0,1\}^n}$, namely a subset of the set of all computational basis states of $n+1$ qubits that end on 0, with $C \neq \emptyset$, define

$$|\psi_C\rangle = \frac{1}{\sqrt{|C|}} \sum_{x0\in C} |x0\rangle.$$

For $C = \emptyset$ we take

$$|\psi_\emptyset\rangle = |0\rangle^{\otimes n} \otimes |1\rangle.$$

(Note that the $(n+1)^{st}$ qubit only really matters for $|\psi_{\emptyset}\rangle$.) Our set of interest will be

$$\mathscr{S} := \{|\psi_C\rangle \mid C \subseteq \{|x0\rangle\}_{x\in\{0,1\}^n}\}.$$

This discrete set of 2^{2^n} multi-qubit quantum states now gives rise to a class of p-concepts

$$\mathscr{F}_{\mathscr{S}} = \{f_C : X \to [0,1] \mid \exists C \subseteq \{|x0\rangle\}_{x\in\{0,1\}^n} : f_C(x) = |\langle x|\psi_C\rangle|^2\}.$$

This class has large pseudo-dimension, as described in the following lemma.

Lemma 4 *With the notation introduced above, it holds that* $Pdim(\mathscr{F}_{\mathscr{S}}) \geq 2^n$.

Proof Consider the subset of computational basis states $\{|x0\rangle\}_{x\in\{0,1\}^n}$ and the corresponding threshold values $y_{x0} = \frac{1}{2^n} = \min_{C\subseteq\{|x0\rangle\}_{x\in\{0,1\}^n}} \frac{1}{|C|}$ independently of $x0$. By construction of $\mathscr{S}$ and thus $\mathscr{F}_{\mathscr{S}}$ the following holds:

For any $C \subseteq \{|x0\rangle\}_{x\in\{0,1\}^n}$

$$f_C(x0) = |\langle x0|\psi_C\rangle|^2 = \begin{cases} \frac{1}{|C|} & \text{if } |x0\rangle \in C \\ 0 & \text{else} \end{cases}.$$

In particular, we have

$$f_C(x0) \geq y_{x0} \iff |x0\rangle \in C.$$

Hence, $\text{Pdim}(\mathscr{F}_{\mathscr{S}}) \geq 2^n$, because we have found an example of a set of size 2^n that is pseudo-shattered. □

We now combine this simple observation with Theorem 3, which gives us the following:

Theorem 5 *With the notation introduced above, if* γ *and* δ *are such that each state in* $\mathscr{S}$ *can be generated from the state* $|0\rangle^{\otimes(n+1)}$ *by some circuit of size* γ *and depth* δ*, then*

$$2^n \leq \mathscr{O}\left(\delta \cdot 2^4 \cdot \gamma^2 \log \gamma\right)$$

Proof Under the assumption of the Theorem we can conclude $\mathscr{F}_{\mathscr{S}} \subseteq \mathscr{F}_{\delta,\gamma}$. Now combine the lower bound of Lemma 4 with the upper bound from Theorem 3. □

Corollary 2 *There exists a* $C \subseteq \{|x0\rangle\}_{x\in\{0,1\}^n}$ *such that* $|\psi_C\rangle = \frac{1}{\sqrt{|C|}} \sum_{|x0\rangle\in C} |x0\rangle$ *cannot be implemented by a quantum circuit of 2-qubit unitaries with subexponential (in n) size or depth.*

Note that any set of functions which pseudo-shatters a set of size 2^n has to have at least 2^{2^n} elements. Hence, the large size of the set C is an automatic consequence of our line of reasoning.

Remark 3 We note that a set of n-qubit states with cardinality doubly exponential in n s.t. at least one of them needs an exponential number of gates (up to logarithmic factors) to be implemented can also be obtained with more standard reasoning. Namely, it is well known that there are n-qubit states the approximation of which up to trace-distance ε requires $\Omega\left(\frac{2^n \log\left(\frac{1}{\varepsilon}\right)}{\log(n)}\right)$ unitary gates (see Nielsen and Chuang 2010, chap. 4.5.4). So if we pick a $\frac{1}{2}$-net of size $\mathscr{O}\left(2^{2^n}\right)$ for the set of pure n-qubit quantum states, this will have the desired properties.

We sketch another way of using our pseudo-dimension bound to study the gate complexity of state preparation and which might lead to a smaller set of candidates. Given n-qubit pure states $|\psi_1\rangle, \ldots, |\psi_m\rangle$ and efficiently implementable (i.e., with polynomially many 2-qubit unitary gates arranged in polynomially many layers) unitaries $U_1, \ldots, U_k$, one can study the set of states $\{U_i|\psi_j\rangle\}_{1\leq i\leq k, 1\leq j\leq m}$.

If an exponential (in n) pseudo-dimension lower bound can be established for

$$\{f : X \to [0,1] \mid \exists 1 \leq i \leq k, 1 \leq j \leq m : f(x) = |\langle x|U_i|\psi_j\rangle|^2\},$$

then, since every U_i is efficiently implementable, one can conclude that at least one among the states $|\psi_j\rangle$ is not efficiently implementable.

The advantage of such a pseudo-dimension-based reasoning would be that m need not be doubly exponential in n, since we can compensate for this in k. This realization can already be used to reduce the size of the set of candidate states given in Corollary 2. However, we have not yet been able to identify sufficiently many efficiently implementable unitaries to reduce the size below doubly exponential. Nevertheless, there is likely room for improvement in applying our method to the gate complexity of quantum state preparation.

4.2 Learnability of quantum circuits

We now use our pseudo-dimension bounds to study learnability. Specifically, we use the pseudo-dimension bound for the case of variable inputs (Section 3.3) combined with the generalization to quantum operations (Section 3.4). We proceed quite similarly to Aaronson (2007).

The learning problem which we want to study is the following: Let μ be a probability measure on $(X \times Y) \times [0,1]$, unknown to the learner. Let $S = \{((x^{(i)}, y^{(i)}), p^{(i)})\}_{i=1}^m$ be corresponding training data drawn i.i.d. according to μ. A learner must, upon input of training data S, size $\Gamma \in \mathbb{N}$, depth $\Delta \in \mathbb{N}$, confidence $\delta \in [0,1)$, accuracy, $\varepsilon \in [0,1)$ and error margin $\beta \in (0,1)$, output a hypothesis quantum circuit $\mathscr{N}$ of size Γ and depth Δ consisting of two-qudit

operations such that, with probability $\geq 1-\delta$ with regard to the choice of training data,

$$\mathbb{P}_{((x,y),p)\sim\mu}\left[|f_{\mathcal{N}}(x,y)-p|>\beta\right]\leq\varepsilon + \inf_{\mathcal{M}}\mathbb{P}_{(x,p)\sim\mu}\left[|f_{\mathcal{M}}(x,y)-p|>\beta\right],$$

where the infimum runs over all quantum circuits $\mathcal{M}$ of size Γ and depth Δ. Here, $f_{\mathcal{N}}$ denotes the function $f_{\mathcal{N}}(x,y)=\langle x|T_{\mathcal{N}}(|y\rangle\langle y|)|x\rangle$ and $f_{\mathcal{M}}$ is defined analogously, similarly to Section 3.3.

We use our pseudo-dimension bound in order to upper bound the size of the training data sufficient for solving this task. More precisely, we make use of sample complexity upper bounds from the fat-shattering dimension as proved in Anthony and Bartlett (2000) and Bartlett and Long (1998), together with the fact that the fat-shattering dimension is upper-bounded by the pseudo-dimension.

First we restrict our scope to the "realizable" scenario, i.e., we will assume the probability measure to be of the form

$$\mu((x,y),p)=\begin{cases}\mu_1(x,y) & \text{if } p=f_{\mathcal{N}_*}(x,y)\\ 0 & \text{else}\end{cases}$$

for some quantum circuit $\mathcal{N}_*$ of size Γ and depth Δ. This will in particular imply that for quantum circuits $\mathcal{M}$ of size Γ and depth Δ

$$\inf_{\mathcal{M}}\mathbb{P}_{((x,y),p)\sim\mu}\left[|f_{\mathcal{M}}(x,y)-p|>\beta\right]=0.$$

Colloquially, realizability means that there exists a set of "correct" parameters Γ and Δ and these are known to the learner, i.e., training samples are promised to be drawn from circuits of size Γ and depth Δ.

We will focus on a proper learning scenario, i.e., we will assume the unknown target circuit to be in some (known) class, namely the class of circuits whose size and depth satisfy certain polynomial bounds, and require the learner to output an element of that same class as hypothesis.

We will make use of the following classical result:

Theorem 6 (Anthony and Bartlett 2000, Corollary 3.3) *Let X be an input space, let $\mathcal{F}\subseteq[0,1]^X$. Let D be a probability measure on X, let $f_*\in\mathcal{F}$. Let $\delta,\varepsilon,\alpha,\beta\in(0,1)$ with $\beta>\alpha$. Let $\mathcal{S}=\{x_1,\ldots,x_m\}$ be a set of m samples drawn i.i.d. according to D. Let $h\in\mathcal{F}$ be such that $|h(x_i)-f_*(x_i)|\leq\alpha$ for all $1\leq i\leq m$.*

Then, a sample size

$$m=\mathcal{O}\left(\frac{1}{\varepsilon}\left(fat_{\mathcal{F}}\left(\frac{\beta-\alpha}{8}\right)\log^2\left(\frac{fat_{\mathcal{F}}\left(\frac{\beta-\alpha}{8}\right)}{(\beta-\alpha)\varepsilon}\right)+\log\frac{1}{\delta}\right)\right)$$

suffices to guarantee that, with probability $\geq 1-\delta$ with regard to the choice of training data $\mathcal{S}$,

$$\mathbb{P}_{x\sim D}[|h(x)-f_*(x)|>\beta]\leq\varepsilon.$$

In our setting, this result implies:

Corollary 3 *Let $\mathcal{N}_*$ be a quantum circuit of quantum operations with size Γ and depth Δ. Let μ be probability measure on $X\times Y$ unknown to the learner. Let*

$$S=\{((x^{(i)},y^{(i)}),f_{\mathcal{N}_*}(x^{(i)},y^{(i)})\}_{i=1}^m$$

be corresponding training data drawn i.i.d. according to μ. Let $\delta,\varepsilon,\alpha,\beta\in(0,1)$. Then, training data of size $m=\mathcal{O}\left(\frac{1}{\varepsilon}\left(\Delta d^8\Gamma^2\log(\Gamma)\log^2\left(\frac{\Delta d^8\Gamma^2\log(\Gamma)}{(\beta-\alpha)\varepsilon}\right)+\log\frac{1}{\delta}\right)\right)$ suffice to guarantee that, with probability $\geq 1-\delta$ with regard to choice of the training data, any quantum circuit $\mathcal{N}$ of size Γ and depth Δ that satisfies

$$|f_{\mathcal{N}}(x_i,y_i)-f_{\mathcal{N}_*}(x_i,y_i)|\leq\alpha\quad\forall 1\leq i\leq m$$

also satisfies

$$\mathbb{P}_{(x,y)\sim\mu}[|f_{\mathcal{N}}(x,y)-f_{\mathcal{N}_*}(x,y)|>\beta]\leq\varepsilon.$$

Proof Combine Theorem 6 with Theorem 3 (more precisely, with its version for variable input states, which can be proved for operations analogously to the reasoning in Section 3.3) and use that the fat-shattering dimension is always upper-bounded by the pseudo-dimension. □

Note that in particular, this implies that for the class of circuits of quantum operations with polynomial size and depth in the number of qudits, a hypothesis that performs well on training data will also perform well in a probably approximately correct sense.

Next, we want to discuss briefly how our result compares to the work (Aaronson 2007) on the learnability of quantum states. There, it is shown that quantum states can be PAC-learned with a sample complexity that depends linearly on the number of qubits and (among other dependencies) polynomially on $\frac{1}{\varepsilon}$, where ε denotes the desired accuracy. However, this result does not imply learnability of quantum channels with a sample complexity that depends polynomially on the number of qubits. This observation is already stated in Aaronson (2007), and we provide an alternate, intuitive explanation for why the result on states does not directly apply to operations.

One can straightforwardly apply the result of Aaronson (2007) to learn the Choi-Jamiolkowski state of a quantum channel. One can then compute measurement probabilities of output states of a channel T acting on n-qubit states, using its Choi-Jamiolkowski state τ. For this we must make use of the formula

$$\mathrm{tr}[ET(\rho)]=2^n\mathrm{tr}[\tau(E\otimes\rho^T)].$$

Here, we see that any error on the side of the Choi-Jamiolkowski state will be multiplied by a factor exponential in n, and thus in this case the overall n-dependence of the sample complexity bound from Aaronson (2007) becomes exponential via the accuracy-dependence.

This motivates our study of learnability of a restricted class of quantum operations. Finding such operations for which process tomography is possible was left as an open problem in Aaronson (2007). Our answer to this question is that a PAC-version of quantum process tomography is possible when we restrict our scope to operations that can be implemented by quantum circuits of depth and size polynomial in the number of qudits. However, note that this is subject to a realizability assumption: the learner must known in advance a polynomial bound on the size and depth of the circuit. We show that imposing the operations be efficiently implementable automatically reduces the information-theoretic complexity of learning, requiring only a modest number of training examples. We do not make any statement about the computational complexity of this learning task; this remains an open problem.

How can this probably approximately correct version of quantum process tomography be put to use? Given polynomially many uses of a black box implementing an unknown quantum operation of polynomial size and depth, one can exhibit a circuit of two-qudit quantum operations that approximates the unknown channel. In other words, we obtain a classical description of an approximate copy of the channel.

5 Open problems

Finally, in this section we discuss future directions and possible generalizations of our results.

Two natural parameters of a circuit, depth and size, appear polynomially in the pseudo-dimension upper bounds. Notably, these bounds are independent of the number of qudits in the circuit. Are our upper bounds tight in their dependence on size and depth? Can similar techniques produce pseudo-dimension lower bounds? For example, by considering a single 2-qudit unitary it is relatively straightforward to see that the pseudo-dimension of a circuit is $\geq \Omega(d)$. Can we close the gap in dimension-dependence between this linear lower bound and our quartic upper bound?

Our application of pseudo-dimension for lower bounds on the gate complexity of state preparation complements known methods (described, e.g., in Nielsen and Chuang 2010), based on counting dimensions or covering arguments. We exhibit a class of states of size 2^{2^n}, for which at least one has exponential gate complexity of state preparation. Can we exploit this new technique to exhibit a smaller set of states? Perhaps the most exciting application of pseudo-dimension bounds could be provable lower bounds on the gate complexity of state preparation, if the reasoning in Section 4.1 is sharpened or the tools are developed further.

If circuit depth and size are known in advance, one can information-efficiently learn the circuit. If the learner receives training data generated by an approximation of the circuit, does the result still hold? Can the realizability assumption be relaxed?

Does "pretty-good circuit tomography" have applications? On the theory side, this might involve exploiting the learning process as an approximate copy-machine for quantum circuits. Of interest for both theory and experiment is whether circuits can be learned with a reasonable amount of computation. One can imagine progress on this question for process tomography similar to that for state tomography; demonstrating a class of states for which learning is computationally efficient in Rocchetto (2017) made it possible to learn physically interesting states in a laboratory in Rocchetto et al. (2019). An efficiency improvement in the process tomography case might also have experimental ramifications.

Acknowledgments M.C.C. and I.D. thank Michael Wolf for suggesting this problem and both Michael Wolf and Yifan Jia for insightful discussions. Also, M.C.C. and I.D. thank Scott Aaronson, Srinivasan Arunachalam and Andrea Rocchetto for their valuable feedback on an earlier version of this paper. Finally, M.C.C. and I.D. thank the reviewers for their helpful suggestions.

M.C.C. gratefully acknowledges support from the TopMath Graduate Center of the TUM Graduate School at the Technische Universität München, Germany, and from the TopMath Program at the Elite Network of Bavaria. M.C.C. is supported by a doctoral scholarship of the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

I.D. gratefully acknowledges that this material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE 1656518, and by the German Academic Exchange Service (DAAD) under Grant No. 57381410. Any conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the aforementioned institutions.

Funding Open Access funding enabled and organized by Projekt DEAL.

Appendix

Here, we prove Theorem 3, namely that $\mathrm{Pdim}(\mathscr{F}_{\delta,\gamma}) \leq \mathscr{O}(\delta \cdot d^4 \cdot \gamma^2 \log \gamma)$.

Proof (Theorem 3)

We rely upon Lemma 2. Let $\{(x_i, y_i)\}_{i=1}^m \subseteq X \times \mathbb{R}$ be such that for every $C \subseteq \{1, \ldots, m\}$, there exists $f_C \in \mathscr{F}_{\delta,\gamma}$ such that $f_C(x_i) - y_i \geq 0$ if and only if $i \in C$.

By Lemma 2, there exists a set of polynomials $\mathscr{P}_{\delta,\gamma}$ in $2\gamma d^4 + 2d^n$ real variables such that $|\mathscr{P}_{\delta,\gamma}| \leq \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^\delta$ and such that for every $C \subseteq \{1, \ldots, m\}$, there exists a $p_C \in \mathscr{P}_{\delta,\gamma}$ and an assignment Ξ_C to the first $2\gamma d^4$ variables of p_C such that $p_C(\Xi_C, x_i) - y_i \geq 0$ if and only if $i \in C$.

In particular, this implies (using the "moreover"-part of Lemma 2) that the set $\mathscr{P} = \{p(\cdot, x_i) - y_i\}_{i=1}^m \mid p \in \mathscr{P}_{\delta,\gamma}\}$ is a set of $m \cdot |\mathscr{P}_{\delta,\gamma}| \leq m \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}\,(n!)^\delta$ polynomials of degree $\leq 2\gamma$ in $2\gamma d^4$ real variables that has at least 2^m different consistent sign assignments. So by Corollary 1, we have

$$2^m \leq \left(\frac{8e \cdot 2\gamma \cdot m}{2\gamma d^4} \cdot \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}\,(n!)^\delta\right)^{2\gamma d^4}.$$

Taking logarithms yields

$$m \leq 2\gamma d^4 \left(\log(16e \cdot \gamma) + \log\left(\frac{m}{2\gamma d^4} \cdot \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}\,(n!)^\delta\right)\right).$$

Repeating the argument in the proof of Theorem 2, we distinguish cases and observe that in both cases,

$$m \leq 8d^4 \cdot \gamma \cdot \log\left(16e\gamma \cdot \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}\,(n!)^\delta\right).$$

Expanding the logarithm and using Stirling's formula up to two terms, we have

$$\begin{aligned}
&\log\left(16e\gamma \cdot \frac{\gamma!\,\delta^{\gamma-\delta}}{(\gamma-\delta)!}\,(n!)^\delta\right)\\
&= \frac{1}{\ln 2}\,(4\ln 2 + 1 + \ln\gamma + [n \cdot \delta \ln n - n \cdot \delta + \mathscr{O}(\ln n)\\
&\quad + \gamma \ln\gamma - \gamma + \mathscr{O}(\ln\gamma) - (\gamma-\delta)\ln(\gamma-\delta)\\
&\quad + (\gamma - \delta) + \mathscr{O}(\ln(\gamma-\delta)) + (\gamma-\delta)\ln\delta])\\
&\leq \frac{1}{\ln 2}\,(4\ln 2 + 1 + \ln\gamma + [2\gamma \cdot \delta(\ln(2\gamma) - 1)\\
&\quad + \gamma\ln\gamma - (\gamma-\delta)\ln(\gamma-\delta) - \delta + (\gamma-\delta)\ln\delta])\\
&= \mathscr{O}(\gamma \cdot \delta \log\gamma).
\end{aligned}$$

We use the fact that $n \leq 2\gamma$ (because we assume that each qudit is acted upon by at least one gate) in the second step, and note that because $\gamma \geq \delta$, the asymptotic behavior of all of the above terms are subsumed by the first term in the bracket. We have also confirmed that the $\log(16e\gamma)$ term above may be neglected. Thus, by the definition of the pseudo-dimension we conclude $\mathrm{Pdim}(\mathscr{F}_{\mathscr{N}}) \leq \mathscr{O}(\delta \cdot d^4 \cdot \gamma^2 \log\gamma)$. □

References

Aaronson S (2007) The learnability of quantum states. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 463(2088):3089–3114. https://doi.org/10.1098/rspa.2007.0113

Aaronson S (2016) The complexity of quantum states and transformations: From quantum money to black holes. Electronic Colloquium on Computational Complexity (ECCC) 23:109

Aaronson S (2018) Shadow tomography of quantum states. http://dl.acm.org/ft_gateway.cfm?id=3188802&type=pdf

Aaronson S, Chen X, Hazan E, Kale S (2018) Online learning of quantum states. http://dl.acm.org/ft_gateway.cfm?id=3327572&type=pdf

Alon N, Ben-David S, Cesa-Bianchi N, Haussler D (1997) Scale-sensitive dimensions, uniform convergence, and learnability. J ACM 44(4):615–631. https://doi.org/10.1145/263867.263927

Anthony M, Bartlett PL (2000) Function learning from interpolation. Comb Probab Comput 9(3):213–225. https://doi.org/10.1017/S0963548300004247

Arunachalam S, de Wolf R (2017) Guest column: A survey of quantum learning theory. SIGACT News 48 , https://pure.uva.nl/ws/files/25255496/p41_arunachalam.pdf

Bartlett PL, Long PM (1998) Prediction, learning, uniform convergence, and scale-sensitive dimensions. J Comput Sys Sci 56(2):174–190. https://doi.org/10.1006/jcss.1997.1557

Bartlett PL, Mendelson S (2002) Rademacher and gaussian complexities: Risk bounds and structural results. J Mach Learn Res 3(Nov):463–482. http://www.jmlr.org/papers/volume3/bartlett02a/bartlett02a.pdf

Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1989) Learnability and the vapnik-chervonenkis dimension. J ACM 36(4):929–965. https://doi.org/10.1145/76359.76371

Cheng HC, Hsieh MH, Yeh PC (2016) The learnability of unknown quantum measurements. Quantum Information & Computation 16(7-8):615–656

Chung KM, Lin HH (2018) Sample efficient algorithms for learning quantum channels in pac model and the approximate state discrimination problem. arXiv:1810.10938

Goldberg PW, Jerrum MR (1995) Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. Mach Learn 18(2-3):131–148. https://doi.org/10.1007/BF00993408

Hanneke S (2016) The optimal sample complexity of pac learning. J Mach Learn Res 17(1):1319–1333. http://dl.acm.org/ft_gateway.cfm?id=2946683&type=pdf

Heinosaari T, Ziman M (2013) The mathematical language of quantum theory: From uncertainty to entanglement. Cambridge University Press, Cambridge

Karpinski M, Macintyre A (1997) Polynomial bounds for vc dimension of sigmoidal and general pfaffian neural networks. J Comput Sys Sci 54(1):169–176. https://doi.org/10.1006/jcss.1997.1477

Kiani BT, Lloyd S, Maity R (2020) Learning unitaries by gradient descent. arXiv:2001.11897

Koiran P (1996) VC dimension in circuit complexity. In: Cai JY, Homer S (eds) Proceedings, Eleventh annual ieee conference on

computational complexity. IEEE Computer Society Press, Los Alamitos, pp 81–85. https://doi.org/10.1109/CCC.1996.507671

Nielsen MA, Chuang IL (2010) Quantum computation and quantum information. Cambridge University Press, Cambridge and New York

Pollard D (1984) Convergence of stochastic processes. Springer Series in Statistics. Springer, New York. https://doi.org/10.1007/978-1-4612-5254-2

Rocchetto A (2017) Stabiliser states are efficiently pac-learnable. Quantum Information and Computation, 18

Rocchetto A, Aaronson S, Severini S, Carvacho G, Poderini D, Agresti I, Bentivegna M, Sciarrino F (2019) Experimental learning of quantum states. Science Advances 5(3), https://doi.org/10.1126/sciadv.aau1946

Torlai G, Mazzola G, Carrasquilla J, Troyer M, Melko R, Carleo G (2018) Neural-network quantum state tomography. Nat Phy 14(5):447–450. https://doi.org/10.1038/s41567-018-0048-5, https://www.nature.com/articles/s41567-018-0048-5.pdf

Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142. https://doi.org/10.1145/1968.1972

Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability & Its Applications 16(2):264–280. https://doi.org/10.1137/1116025

Warren HE (1968) Lower bounds for approximation by nonlinear manifolds. Trans Am Math Soc 133(1):167. https://doi.org/10.2307/1994937

A.2 Necessary criteria for Markovian divisibility of linear maps

Necessary criteria for Markovian divisibility of linear maps

Matthias C. Caro and Benedikt R. Graswald

Markovian evolutions, in which the future is independent of the past, given the present, have been studied intensely, both in classical probability theory and in quantum theory. In particular, the generators of continuous one-parameter semigroups have been identified: Transition rate matrices generate semigroups of stochastic matrices, Lindblad generators generate semigroups of quantum channels. These semigroups and their generators describe time-homogeneous Markovian evolutions of classical and quantum systems, respectively. In this work, we investigate a notion of divisibility describing time-inhomogeneous Markovian evolutions.

After an introduction, in which we motivate our problem, give an overview of our results, and discuss related work, we recall some well known notions from quantum information theory in Section II. Next, in Subsection III.A, we define Markovian divisibility (Definition III.1) and infinitesimal Markovian divisibility (Definition III.2) of a linear map with respect to a general (compact and convex) set of generators, the two central notions studied in this article. After discussing some basic properties related to these definitions, we note that, when taking as our set of generators the Lindblad generators, we recover the notion of infinitesimal Markovian divisibility of quantum channels introduced in [10]. In Subsection III.B, we recall from [10] that such infinitesimal Markovian divisible quantum channels have nonnegative determinant, and that in the qubit case, they have been fully characterized.

Section IV contains our main results. Throughout, our goal is to prove that an (infinitesimal) Markovian divisible map $T \in \mathcal{B}(\mathbb{C}^d)$ satisfies an inequality of the form

$$|\det(T)| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T)\right)^{p} , \tag{A.2.1}$$

for suitably chosen d-dependent parameters $p \in \mathbb{R}$ and $k \in \{1, 2, \ldots, d\}$. In Subsection IV.A, we present a general proof strategy for deriving such an inequality from properties of the generators. Namely, we first exploit submultiplicativity of products of largest singular values to show that, if T_1 and T_2 satisfy Eq. (A.2.1), then so does their product $T_1 T_2$ (Lemma IV.1). Next, with the case of infinitesimal Markovian divisibility in mind, we combine this first insight with Trotterization to establish an analogous result for exponentials of generators: If e^{G_1} and e^{G_2} satisfy Eq. (A.2.1), then so does $e^{G_1+G_2}$ (Lemma IV.2). Using a majorization inequality for the singular values of a matrix exponential, we then prove in Lemma IV.4 that a sufficient condition for e^G to fulfill the singular value inequality Eq. (A.2.1) is that G satisfies the eigenvalue inequality

$$\mathrm{tr}[G + G^*] - p \sum_{i=1}^{k} \lambda_i^{\uparrow}(G + G^*) \leq 0 \, . \tag{A.2.2}$$

Together with continuity of the determinant, Lemmas IV.1 and IV.4 lead to our first main result (Theorem IV.5): If every admissible generator G satisfies Eq. (A.2.2), then any map T that is Markovian divisible with respect to that set of generators satisfies Eq. (A.2.1). In the case of

infinitesimal divisibility, we employ Lemma IV.2 to show that it suffices to have Eq. (A.2.2) for positive multiples of extreme points of the convex and compact set of generators to guarantee Eq. (A.2.1) (Corollary IV.6).
In Subsection IV.B, we apply this strategy for the case of infinitesimal Markovian divisible quantum channels. Namely, in Lemmas IV.7, IV.14, and Proposition IV.21, we show that Lindblad generators on qudits satisfy Eq. (A.2.2) for the parameter settings $(p,k) \in \{(d/2,1),(1,\lfloor 2d - 2\sqrt{2d}+1\rfloor)\} \cup \{(2d/k+2\sqrt{d}+1,k) \mid 1 \leq k \leq d^2\}$. Consequently, according to the results of Subsection IV.A, infinitesimal Markovian divisible quantum channels satisfy Eq. (A.2.1) with the same choices of p and k. Thus, we have proved necessary criteria for the infinitesimal Markovian divisibility of quantum channels in terms of an upper bound on the determinant via smallest singular values. Example IV.11 describes an analytical application of these criteria in identifying new examples of not infinitesimal Markovian divisible quantum channels. Moreover, we argue in Examples IV.12 and IV.17 that our parameter choices (p,k) are close to optimal in general, prove a small improvement in the Appendix, and discuss strengthenings of our results for normal Lindbladians in Proposition IV.13 and Remark IV.15.
To demonstrate that the necessary criteria obtained in Subsection IV.B are indeed quantum features, we show in Subsection IV.C that no non-trivial necessary criteria of the same form can hold in the classical case, with arbitrary transition rate matrices as generators. We do so by studying a concrete example (Example IV.24). However, as we demonstrate in Lemma IV.26 and Corollary IV.27, after appropriately restricting the set of generators to a subset of transition rate matrices, the proof strategy from Subsection IV.A can be applied successfully. With Section V., we conclude the article with a short discussion of our results and some open questions, including a concrete conjecture (Conjecture IV.19).
I was significantly involved in finding the ideas and carrying out the scientific work of all parts of this article. The idea for this project was motivated by discussions between my doctoral advisor, Michael M. Wolf, and myself. In these discussions, also the first idea for the general proof strategy was developed. All other parts of the scientific work for this article were completed by Benedikt R. Graswald and myself. I was in charge of writing the article, with the exception of Corollary IV.10., Example IV.11., Proposition IV.13., and the Appendix.

Permission to include:

Matthias C. Caro and Benedikt R. Graswald.
Necessary criteria for Markovian divisibility of linear maps.
Journal of Mathematical Physics 62, 042203 (2021). https://doi.org/10.1063/5.0031760.

AIP Publishing LLC

Your Window to Possible

Permission to Reuse Content

REUSING AIP PUBLISHING CONTENT

Permission from AIP Publishing is required to:

- republish content (e.g., excerpts, figures, tables) if you are not the author
- modify, adapt, or redraw materials for another publication
- systematically reproduce content
- store or distribute content electronically
- copy content for promotional purposes

To request permission to reuse AIP Publishing content, use RightsLink® for the fastest response or contact AIP Publishing directly at rights@aip.org (mailto:rights@aip.org) and we will respond within one week:

For RightsLink, use Scitation to access the article you wish to license, and click on the Reprints and Permissions link under the TOOLS tab. (For assistance click the “Help” button in the top right corner of the RightsLink page.)

To send a permission request to rights@aip.org (mailto:rights@aip.org), please include the following:

- Citation information for the article containing the material you wish to reuse
- A description of the material you wish to reuse, including figure and/or table numbers
- The title, authors, name of the publisher, and expected publication date of the new work
- The format(s) the new work will appear in (e.g., print, electronic, CD-ROM)
- How the new work will be distributed and whether it will be offered for sale

Authors do **not** need permission from AIP Publishing to:

- quote from a publication (please include the material in quotation marks and provide the customary acknowledgment of the source)
- reuse any materials that are licensed under a Creative Commons CC BY license (please format your credit line: “Author names, Journal Titles, Vol.#, Article ID#, Year of Publication; licensed under a Creative Commons Attribution (CC BY) license.”)
- reuse your own AIP Publishing article in your thesis or dissertation (please format your credit line: “Reproduced from [FULL CITATION], with the permission of AIP Publishing”)
- reuse content that appears in an AIP Publishing journal for republication in another AIP Publishing journal (please format your credit line: “Reproduced from [FULL CITATION], with the permission of AIP Publishing”)
- make multiple copies of articles–although you must contact the Copyright Clearance Center (CCC) at www.copyright.com (http://www.copyright.com/) to do this

R ING CONTENT PUBLISHED BY OTHERS

content.)

Reuse of Previously Published Material Form (pdf (https://publishing.aip.org/wp-content/uploads/AIP_Permission_Form-1.pdf))

Unless the publisher requires a specific credit line, please format yours like this:

Reproduced with permission from J. Org. Chem. 63, 99 (1998). Copyright 1998, American Chemical Society.

You do not need permission to reuse material in the public domain, but you should still include an appropriate credit line which cites the original source.

Necessary criteria for Markovian divisibility of linear maps

Cite as: J. Math. Phys. **62**, 042203 (2021); https://doi.org/10.1063/5.0031760
Submitted: 03 October 2020 . Accepted: 24 March 2021 . Published Online: 13 April 2021

Matthias C. Caro, and Benedikt R. Graswald

Journal of Mathematical Physics ARTICLE scitation.org/journal/jmp

Necessary criteria for Markovian divisibility of linear maps

Cite as: J. Math. Phys. 62, 042203 (2021); doi: 10.1063/5.0031760
Submitted: 3 October 2020 • Accepted: 24 March 2021 •
Published Online: 13 April 2021

Matthias C. Caro[1,2,a)] and Benedikt R. Graswald[1,b)]

AFFILIATIONS

[1]Technical University of Munich, Department of Mathematics, Boltzmannstraße 3, 85748 Garching bei München, Germany
[2]Munich Center for Quantum Science and Technology (MCQST), Munich, Germany

[a)]**Author to whom correspondence should be addressed:** caro@ma.tum.de. **URL:** https://sites.google.com/view/matthiasccaro
[b)]graswabe@ma.tum.de. **URL:** https://www-m7.ma.tum.de/bin/view/Analysis/BenediktGraswald

ABSTRACT

We describe how to extend the notion of infinitesimal Markovian divisibility from quantum channels to general linear maps and compact and convex sets of generators. We give a general approach toward proving necessary criteria for (infinitesimal) Markovian divisibility. With it, we prove two necessary criteria for infinitesimal divisibility of quantum channels in any finite dimension d: an upper bound on the determinant in terms of a $\Theta(d)$-power of the smallest singular value and in terms of a product of $\Theta(d)$ smallest singular values. These allow us to analytically construct, in any given dimension, a set of channels that contains provably non-infinitesimal Markovian divisible ones. Moreover, we show that, in general, no such non-trivial criteria can be derived for the classical counterpart of this scenario.

I. INTRODUCTION

References 1 and 2 made an important step toward understanding the connection between master equations and the framework of quantum channels for describing quantum evolutions by characterizing the generators, which give rise to semigroups of quantum channels via the corresponding (time-independent) master equation. The converse question, i.e., the problem of characterizing those quantum channels that can arise from the solution of a (possibly time-dependent) Lindblad master equation, is, however, still awaiting an answer.

Endeavors toward a resolution of this problem have given rise to different notions of (non-) Markovianity for quantum evolutions. One line of research is based on connecting Markovianity to certain divisibility properties of quantum evolutions, particularly to the possibility of dividing the evolution into infinitesimal pieces. While this gives an intuitively plausible notion of time-dependent quantum Markovianity and some structural properties can be established on its basis, it has so far not given rise to easily verifiable criteria for Markovianity (with a simple exception). Only for evolutions of qubit systems is this notion completely understood. We go beyond this characterization for the two-dimensional case and establish necessary criteria for a quantum channel—or a linear map in general—to be divisible into infinitesimal Markovian pieces. Our criteria take the form of an upper bound on the determinant in terms of the power of a product of smallest singular values.

Our proof strategy is not specific to quantum channels but can be applied to obtain necessary criteria for (infinitesimal) Markovian divisibility of general linear maps with respect to a closed and convex set of generators if the generators satisfy certain spectral properties.

A. Overview of our results

In this work, we study the following question: Given a linear map T and a set of linear maps $\mathcal{G}$, acting on $\mathbb{C}^d$, can T be approximated arbitrarily well by linear maps of the form $\prod_i e^{G_i}$, where $G_i \in \mathcal{G}$? If that is the case, we say that T is *Markovian divisible with respect to the set of generators* $\mathcal{G}$.

We aim toward establishing necessary criteria for Markovian divisibility of the form

$$|\det(T)| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T)\right)^{p},$$

where $k = k(d)$ and $p = p(d)$ depend on the underlying dimension. Proving such criteria becomes tractable by combining multiplicativity of the determinant and sub-/super-multiplicativity of products of largest/smallest singular values with Trotterization.

In Sec. IV A, we describe how to use these properties to reduce the problem of establishing necessary criteria of the above form to a spectral property of the generators. We can summarize our reduction as follows:

Theorem *(Theorem IV.5—informal version). Let $\mathcal{G} \subseteq \mathcal{M}_d$ be a set of generators. Let T be Markovian divisible with respect to $\mathcal{G}$, and suppose that every $G \in \mathcal{G}$ satisfies $\mathrm{Tr}[G + G^*] - p\sum_{i=1}^{k} \lambda_i^{\uparrow}(G + G^*) \leq 0$. Then, $|\det(T)| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T)\right)^{p}$.*

We employ our proof strategy for the physically motivated scenario of *infinitesimal Markovian divisibility*. Here, the objects of interest are linear maps T that, for any $\varepsilon > 0$, can be arbitrarily well approximated by linear maps of the form $\prod_i e^{G_i}$, where $G_i \in \mathcal{G}$ are such that $\left\|e^{G_i} - \mathbb{1}_d\right\| \leq \varepsilon$.

We first study the case in which $\mathcal{G}$ is the set of Lindblad generators seen as linear maps on $d \times d$-matrices, i.e., we consider those generators that give rise to semigroups of quantum channels. With this choice, the notion of infinitesimal Markovian divisibility of a linear map T on $d \times d$-matrices becomes that of infinitesimal Markovian divisibility of quantum channels introduced in Ref. 3.

We prove necessary criteria for infinitesimal Markovian divisibility of quantum channels in any finite dimension. Specifically, for an infinitesimal Markovian divisible quantum channel T on $d \times d$-matrices, we show in Corollaries IV.9 and IV.16 that

$$|\det(T)| \leq \left(s_1^{\uparrow}(T)\right)^{\frac{d}{2}} \text{ and } |\det(T)| \leq \prod_{i=1}^{\lfloor 2d-2\sqrt{2d}+1 \rfloor} s_i^{\uparrow}(T).$$

Moreover, we give explicit examples (Examples IV.12 and IV.17) of infinitesimal divisible channels from which we can conclude that the d-dependence of the exponent (in the first bound) and of the number of singular value factors (in the second bound) is close to optimal, respectively.

We also describe how to interpolate between these bounds in Corollary IV.21 and obtain that for an infinitesimal divisible quantum channel T acting on $d \times d$-matrices,

$$|\det(T)| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T)\right)^{\frac{2d}{k+2\sqrt{k}+1}} \quad \text{for } 1 \leq k \leq d^2.$$

These criteria allow us to give new examples of provably non-infinitesimal divisible channels in dimensions strictly bigger than 2, which were not recognizable as such previously (Example IV.11).

As a second application of our proof strategy, we take $\mathcal{G}$ to be the set of transition rate matrices of dimension d and thereby study the question of (infinitesimal) Markovian divisibility of stochastic matrices. We first show via an explicit example (Example IV.24) that no necessary criterion of the above form can hold in this scenario when we allow all transition rate matrices as generators. Combined with our results for infinitesimal Markovian divisible quantum channels, this implies that stochastic matrices cannot be embedded into quantum channels while preserving both the singular values and the property of infinitesimal Markovian divisibility at the same time.

If, however, we restrict our set of generators to transition rate matrices whose diagonal elements differ by at most a constant factor, our proof strategy can be applied and yields an upper bound on the determinant in terms of a power of the smallest singular value (Corollary IV.27).

B. Related work

The quantum Markovianity problem, the question of deciding whether a given quantum channel is a member of a quantum dynamical semigroup, was considered from a complexity-theoretic perspective in Ref. 4. Therein, it was shown to be NP-hard and the same is true for the classical counterpart of this problem, with stochastic matrices instead of quantum channels and transition rate matrices instead of Lindblad generators. The computational complexity of a related divisibility problem for stochastic matrices, namely, that of finite divisibility, was studied in Ref. 5. In addition, this divisibility problem turns out to be NP-hard, even NP-complete.

When fixing the system dimension, however, deciding whether a quantum channel is an exponential of a Lindblad generator, in which case it can be called time-independent Markovian because it solves a time-independent Lindblad master equation, becomes feasible. Corresponding necessary and sufficient criteria and an efficient (in the desired precision) algorithmic procedure for this case with a fixed dimension were given in Refs. 4 and 6. These results pertain to time-independent (quantum) Markovianity and cannot be directly applied to the time-dependent case.

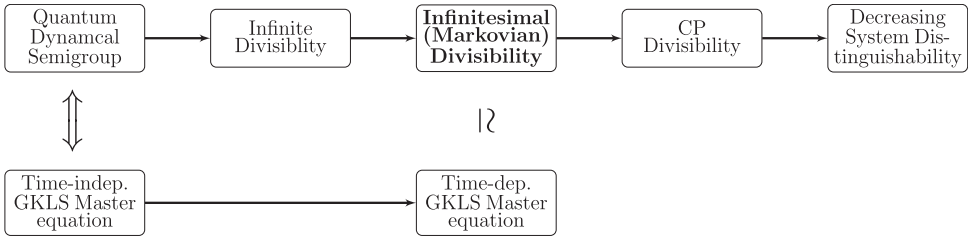


FIG. 1. A depiction of the relations between different notions of divisibility and Markovianity of quantum channels and quantum dynamical maps. A simple arrow indicates that a channel or dynamical map satisfying the condition at the tail also satisfies that at the head. ⇳ indicates the equivalence of two notions. ≃ is used to indicate a correspondence that, to the best of our knowledge, has been rigorously proven only for the qubit case.

Our focus is on infinitesimal Markovian divisible quantum channels. These were introduced and studied in detail for qubit channels by Ref. 3. Therein, it is also observed that every infinitely divisible quantum channel, i.e., every channel that can be written as an nth power of a quantum channel for every $n \in \mathbb{N}$, is infinitesimal divisible. The notion of infinitesimal Markovian divisibility can be seen as corresponding to time-dependent Markovianity, i.e., to solutions of time-dependent Lindblad master equations. Thereby, it offers a route to studying a time-dependent version of the Markovianity problem.

A plethora of different notions of Markovianity for quantum evolutions and relations between them are discussed in several review papers.[7–10] On the one hand, one considers notions of quantum Markovianity based on the divisibility of the evolution, either for quantum channels or for quantum dynamical maps with corresponding propagators. This line of research was initiated by Ref. 3, and Refs. 11 and 12 constitute recent additions to it. In relation to this approach, Ref. 13 proposed a measure of non-Markovianity on the basis of infinitesimal deviations from complete positivity. On the other hand, there are notions and measures of non-Markovianity based on (quantum) information backflow, often formalized in terms of distinguishability measures that are known to be non-increasing under completely positive and trace-preserving maps. This idea was introduced in Ref. 14, and Ref. 9 recently proposed a variant of it.

In Fig. 1, we present only a selected few of these notions and the connections between them.

C. Structure of the paper

Section II introduces basic notions from quantum information that provide our overall framework. In Sec. III, we introduce the core definition of infinitesimal Markovian divisibility in a general setting and discuss prior work in the quantum scenario. Section IV contains our main results: We describe the general proof approach in Subsection IV A and apply it to derive necessary criteria for infinitesimal Markovian divisibility of quantum channels in Subsection IV B. The same type of criterion does not, in general, hold for infinitesimal divisibility of stochastic matrices, only for suitable subsets, as we argue in Subsection IV C. We conclude with some open questions and the references.

II. PRELIMINARIES

We introduce some of the basic notions of quantum information with focus on quantum channels and the corresponding semigroups. The interested reader is referred to Ref. 15 for more details.

Throughout this paper, we denote the set of $d \times d$ complex matrices as $\mathcal{M}_d$ for a dimension $d \in \mathbb{N}$. The identity matrix in $\mathcal{M}_d$ is written as $\mathbb{1}_d$, whereas $\mathrm{id} = \mathrm{id}_{\mathcal{M}_d}$ denotes the identity map on $\mathcal{M}_d$. For $A \in \mathcal{M}_d$, we use $\lambda_i = \lambda_i(A)$ to denote its eigenvalues. If $A \in \mathcal{M}_d$ is Hermitian, we use $\lambda_i^{\downarrow}$ ($\lambda_i^{\uparrow}$) to denote the eigenvalues in decreasing (increasing) order. Similarly, we use the notation $s_i^{\downarrow}$ and $s_i^{\uparrow}$ for singular values. Finally, $\mathrm{Tr}[A]$ will denote the trace of A.

A. Quantum states and channels

A d-level quantum system (for $d \in \mathbb{N}$) is described by a $d \times d$ *density matrix*, i.e., an element of

$$\mathcal{S}\left(\mathbb{C}^d\right) := \{\rho \in \mathcal{M}_d \mid \rho \geq 0,\ \mathrm{Tr}[\rho] = 1\},$$

where $\rho \geq 0$ means that the matrix ρ is positive semidefinite.

Physically admissible transformations of quantum systems are described by *quantum channels* (in the Schrödinger picture), i.e., by elements of

$$\mathcal{T}\left(\mathbb{C}^d, \mathbb{C}^{d'}\right) := \{T : \mathcal{M}_d \to \mathcal{M}_{d'} \mid T \text{is linear, completely positive, and trace} - \text{preserving}\}.$$

Here, we call T *completely positive* iff $T \otimes \mathrm{id}_{\mathcal{M}_n}$ is positivity-preserving for every $n \in \mathbb{N}$. This definition guarantees that a quantum channel maps states to states and that this is still the case when embedding the quantum system of interest into a larger system with trivial evolution on the environmental subsystem.

We will also use the shorthand $\mathcal{T}_d := \mathcal{T}\left(\mathbb{C}^d, \mathbb{C}^d\right)$ for channels with equal input and output dimension.

B. Quantum dynamical semigroups

It is a foundational postulate in quantum theory that the dynamics of a closed quantum system can be described in terms of a Schrödinger equation, which gives rise to a one-parameter group of unitaries. For open quantum systems, we will work with one-parameter semigroups.

Definition II.1 (Continuous dynamical semigroups). A family of linear maps $T_t : \mathcal{M}_d \to \mathcal{M}_d$ with time parameter $t \in \mathbb{R}_+$ is called a dynamical semigroup if $\forall t, s \in \mathbb{R}_+ = [0, \infty) : T_t T_s = T_{t+s}$ *and* $T_0 = Id$. *If in addition, the map $t \mapsto T_t$ is continuous (we are working on finite dimensional spaces, so there is no need to specify the type of continuity here), then the family is called a* continuous dynamical semigroup.

It is well-known that such continuous dynamical semigroups can be represented via a generator, i.e., if $\{T_t\}_{t\geq 0}$ is a continuous dynamical semigroup, then there exists a linear map $L : \mathcal{M}_d \to \mathcal{M}_d$ such that $T_t = e^{tL}$ for all $t \geq 0$.

When requiring such a semigroup to consist of physically admissible evolutions of a quantum system, i.e., of quantum channels, the question arises of what the corresponding generators are. This was answered in the following.

Theorem II.2 *(Generators of quantum dynamical semigroups—GKLS,* Refs. 1 and 2*). A linear map $L : \mathcal{M}_d \to \mathcal{M}_d$ is the generator of a continuous dynamical semigroup of quantum channels if and only if it can be written as*

$$L(\rho) = i[\rho, H] + \sum_j \mathcal{L}_j \rho \mathcal{L}_j^\dagger - \frac{1}{2}\{\mathcal{L}_j^\dagger \mathcal{L}_j, \rho\}, \tag{1}$$

where $H = H^\dagger \in \mathcal{M}_d$ is self-adjoint and $\{\mathcal{L}_j\}_j$ is a set of matrices in $\mathcal{M}_d$. Here, $\{\cdot,\cdot\}$ denotes the anti-commutator.

For such generators, often called *GKLS* or *Lindblad generators*, we refer to the term $i[\cdot, H]$ as the Hamiltonian part and to $\sum_j \mathcal{L}_j \cdot \mathcal{L}_j^\dagger - \frac{1}{2}\{\mathcal{L}_j^\dagger \mathcal{L}_j, \cdot\}$ as the dissipative part with Lindbladians $\{\mathcal{L}_j\}_j$.

We will call a quantum channel *Markovian* if it is an element of a quantum dynamical semigroup.

III. MARKOVIAN DIVISIBILITY

The main motivation for our work is the following problem: Given a quantum channel, decide whether it comes from a (possibly time-dependent) Lindblad master equation. We take two different perspectives on this task to motivate our definitions.

The first perspective is that of differential equations. Specifically, we want to understand which quantum channels can arise as a solution of a time-dependent master equation of the form $\frac{\mathrm{d}}{\mathrm{d}t} T_t = L(t) T_t$, where $L(t)$ is a time-dependent Lindblad generator. More generally, we want to study the possible solutions of a linear ordinary differential equation $\frac{\mathrm{d}}{\mathrm{d}t} T_t = G(t) T_t$, where $t \mapsto G(t) \in \mathcal{G}$, with $\mathcal{G} \subset \mathcal{M}_d$ being a fixed set of generators.

Our second perspective on the problem comes from the semigroup structure of the solutions to time-independent master equations. Specifically, each such equation corresponds to a quantum dynamical semigroup. If we now also want to take into account a possible time-dependence of the generator while still preserving the semigroup structure, we can consider the semigroup generated by all elements of quantum dynamical semigroups. On an intuitive level, the question about solutions of master equations that we asked above now becomes the question of whether a given quantum channel is an element of this semigroup, i.e., we are dealing with the membership problem for this semigroup. Again, we can generalize the question by going from Lindblad generators to general generators.

A. Markovian divisibility with respect to general sets of generators

The two perspectives given above lead us to two slightly different definitions. In the first, we focus on the semigroup structure.

Definition III.1 (Markovian divisibility). Let $\mathcal{G} \subset \mathcal{M}_d$ be a set of matrices, whose elements we call generators. We define the set

$$\mathcal{D}_{\mathcal{G}} := \left\{ T \in \mathcal{M}_d \mid \exists n \in \mathbb{N},\ \textit{generators}\ \{G_i\}_{1\leq i\leq n} \subset \mathcal{G}\ \text{so that}\ \prod_{i=1}^{n} e^{G_i} = T \right\}.$$

We call the closure $\overline{\mathcal{D}}_{\mathcal{G}}$ the set of linear maps that are Markovian divisible with respect to $\mathcal{G}$.

When translating the mathematical motivation of semigroups to a more physical motivation, Definition III.1 can be seen as an approach to the question of which linear maps can be arbitrarily well approximated using alternating exponentials of a fixed set of (*control*) generators.

Now, we give a definition based much on the perspective of differential equations determining the overall evolution on infinitesimal time intervals while keeping the semigroup structure in mind.

Definition III.2 (Infinitesimal Markovian divisibility). Let $\mathcal{G} \subset \mathcal{M}_d$ be a compact and convex set of matrices containing $0 \in \mathcal{M}_d$. We will again refer to the elements of $\mathcal{G}$ as generators. We define the set

$$\mathcal{I}_{\mathcal{G}} := \Big\{ T \in \mathcal{M}_d \mid \forall \varepsilon > 0 \ \exists n \in \mathbb{N}, \ \text{generators } \{G_j\}_{1\le j\le n} \subset \mathcal{G} \\ \text{so that } (i)\ \left\| e^{G_j} - \mathbb{1}_d \right\| \le \varepsilon \ \forall\ j \text{ and } (ii) \prod_{j=1}^{n} e^{G_j} = T \Big\}.$$

We call the closure $\overline{\mathcal{I}}_{\mathcal{G}}$ *the set of linear maps that are infinitesimal* Markovian divisible with respect to $\mathcal{G}$.

Remark III.3. In the definition, we require $\mathcal{G}$ to be compact. This can be assumed without loss of generality. First, closedness can be assumed without loss of generality since for non-closed $\mathcal{G}_0$, we have $\overline{\mathcal{I}}_{\mathcal{G}_0} = \overline{\mathcal{I}}_{\overline{\mathcal{G}_0}}$. Second, boundedness can also be assumed without loss of generality. Specifically, suppose that $\tilde{\mathcal{G}} \subset \mathcal{M}_d$ is an unbounded closed and convex set with $0 \in \tilde{\mathcal{G}}$ and $T \in \mathcal{I}_{\tilde{\mathcal{G}}}$. Then, by definition, $\forall \varepsilon > 0$ $\exists n \in \mathbb{N}$ and $\{G_j\}_{1\le j\le n} \subset \tilde{\mathcal{G}}$ such that $\left\| e^{G_j} - \mathbb{1}_d \right\| \le \varepsilon$ and $\prod_{j=1}^{n} e^{G_j} = T$. By convexity, also $\frac{1}{N} G_j \in \tilde{\mathcal{G}}$ $\forall 1 \le j \le n$ for every $N > 1$. By continuity of the matrix exponential, there exists $N_0 \in \mathbb{N}$ such that $\left\| e^{\frac{1}{N} G_j} - \mathbb{1}_d \right\| \le \varepsilon$ for all $N \ge N_0$. Clearly, we can write $T = \prod_{j=1}^{n} e^{G_j} = \prod_{j=1}^{n} \left(e^{\frac{1}{N} G_j} \right)^N$. Thus, as $\left\| \frac{1}{N} G_j \right\| \to 0$ as $N \to \infty$, we conclude that for every $B > 0$, we have $T \in \mathcal{I}_{\tilde{\mathcal{G}}_{\le B}}$, where $\tilde{\mathcal{G}}_{\le B} := \{G \in \tilde{\mathcal{G}} \mid \|G\| \le B\}$. Hence, we can impose an arbitrary (non-zero) norm bound on our generators without changing the set of infinitesimal Markovian divisible channels.

Therefore, we are justified in using Definition III.2 also for non-compact $\mathcal{G}$ (in particular, Lindblad generators and transition rate matrices).

Remark III.4. By continuity of the matrix exponential, it is easy to see that, if $G \in \mathcal{G}$ implies $\frac{1}{n} G \in \mathcal{G}$ for all $n \in \mathbb{N}$, then $\mathcal{D}_{\mathcal{G}} = \mathcal{I}_{\mathcal{G}}$. This is particularly the case if $\mathcal{G}$ satisfies the assumptions of Definition III.2.

If, however, $\mathcal{G}$ does not have this property, then (i) in the definition of $\mathcal{I}_{\mathcal{G}}$ will, in general, lead to $\mathcal{I}_{\mathcal{G}} \ne \mathcal{D}_{\mathcal{G}}$ (e.g., $\mathcal{I}_{\mathcal{G}}$ could be empty even if $\mathcal{D}_{\mathcal{G}}$ is not).

When specifying $\mathcal{G}$ to be the set of Lindblad generators and thus the linear maps of interest to be quantum channels, Definitions III.1 and III.2 become connected to quantum channels arising from master equations. Studying such channels via a notion of Markovian divisibility into infinitesimal pieces was first proposed in Ref. 3. Next, we discuss some results of that work.

B. Infinitesimal Markovian divisibility of quantum channels

For ease of notation, we will denote by $\mathcal{I}_d$ the set $\mathcal{I}_{\mathcal{G}}$ for the specific choice of $\mathcal{G}$ being the set of Lindblad generators acting on $d \times d$-matrices. Then, the set $\overline{\mathcal{I}}_d$ is the set of *infinitesimal Markovian divisible quantum channels*, as defined in Ref. 3.

When referring to these channels, we will sometimes drop the "Markovian" for convenience. This can also be justified in a rigorous sense (see Theorem 16 in Ref. 3).

While some insight into the structure of infinitesimal Markovian divisible quantum channels has been obtained in Ref. 3, so far, there are no simple-to-check criteria for infinitesimal divisibility for a general dimension d. Such criteria are the main focus of this work.

A straightforward necessary criterion for infinitesimal divisibility is already observed in Ref. 3, namely, we have the following as a direct consequence of multiplicativity and continuity of the determinant:

Proposition III.5. An infinitesimal divisible quantum channel T *satisfies* $\det(T) \ge 0$.

This is, to our knowledge, the only necessary criterion for infinitesimal divisibility known so far that holds in any finite dimension.

For the special case of qubit channels, the set of infinitesimal divisible channels can be explicitly characterized by making use of the Lorentz normal form (the latter is discussed in Ref. 16).

Theorem III.6 (Infinitesimal divisible qubit channels[3]—informal). *Let* $T : \mathcal{M}_2 \to \mathcal{M}_2$ *be a generic qubit channel with the Lorentz normal form* $\begin{pmatrix} 1 & 0 \\ 0 & \Delta \end{pmatrix}$.

T *is infinitesimal Markovian divisible if and only if* $0 \le \det(\Delta) \le s_{\min}^2$, *where* $s_{\min}$ *is the smallest singular value of* Δ.

This characterization serves as one motivation for our results in higher dimensions, which we derive in Subsection IV B.

IV. NECESSARY CRITERIA FOR MARKOVIAN DIVISIBILITY

We now develop necessary criteria for a linear map to be (infinitesimal) Markovian divisible. More precisely, our discussion aims toward establishing inequalities of the form

$$|\det(T)| \le \left(\prod_{i=1}^{k} s_i^{\uparrow}(T) \right)^p. \tag{2}$$

We first present some results for the case of general linear maps and generators and later combine these observations with a more detailed analysis for quantum channels and Lindblad generators and stochastic matrices and transition rate matrices, respectively.

A. General sets of generators

We first observe that if each of two matrices satisfies the desired inequality (2), then so does the product of the matrices.

Lemma IV.1. Let $T_1, T_2 \in \mathcal{M}_d$. Suppose that $1 \le k \le d$ and $p > 0$ such that

$$|\det(T_j)| \le \left(\prod_{i=1}^{k} s_i^{\uparrow}(T_j)\right)^p$$

holds for $j = 1, 2$. Then, also

$$|\det(T_1 T_2)| \le \left(\prod_{i=1}^{k} s_i^{\uparrow}(T_1 T_2)\right)^p.$$

Proof. A well-known majorization inequality for singular values states that

$$\prod_{i=1}^{k} s_i^{\downarrow}(AB) \le \prod_{i=1}^{k} s_i^{\downarrow}(A) s_i^{\downarrow}(B) \tag{3}$$

for any $1 \le k \le n$ for $n \times n$-matrices A, B (see Ref. 17, Theorem 3.3.4). With this, we obtain

$$\begin{aligned}
|\det(T_1 T_2)| &= |\det(T_1)||\det(T_2)| \\
&\le \left(\prod_{i=1}^{k} s_i^{\uparrow}(T_1)\right)^p \left(\prod_{i=1}^{k} s_i^{\uparrow}(T_2)\right)^p \\
&= \left(\frac{|\det(T_1)||\det(T_2)|}{\prod_{i=1}^{d-k} s_i^{\downarrow}(T_1) s_i^{\downarrow}(T_2)}\right)^p \\
&\le \left(\frac{|\det(T_1 T_2)|}{\prod_{i=1}^{d-k} s_i^{\downarrow}(T_1 T_2)}\right)^p \\
&= \left(\prod_{i=1}^{k} s_i^{\uparrow}(T_1 T_2)\right)^p
\end{aligned}$$

as claimed. Here, the first inequality is that, by assumption, the following step uses $|\det(T_i)| = \prod_{j=1}^{d} s_j^{\downarrow}(T_i)$, the second inequality is due to Eq. (3), and the last step uses $|\det(T_1 T_2)| = \prod_{j=1}^{d} s_j^{\downarrow}(T_1 T_2)$. □

This means that, when trying to establish an inequality of the form (2), if T is a finite product, it suffices to consider the single factors separately.

Now we show that, once we have our desired inequality (2) for non-negative multiples of two separate generators, the exponential of the sum of these two generators also satisfies the inequality. This observation will be particularly useful in our analysis of Lindblad generators.

Lemma IV.2. Let $G_1, G_2 \in \mathcal{M}_d$. Suppose that $1 \le k \le d$ and $p > 0$ are such that

$$|\det\left(e^{\frac{G_j}{n}}\right)| \le \left(\prod_{i=1}^{k} s_i^{\uparrow}\left(e^{\frac{G_j}{n}}\right)\right)^p$$

holds for all $n \in \mathbb{N}$ and $j = 1, 2$. Then, also

$$|\det(e^{G_1+G_2})| \le \left(\prod_{i=1}^{k} s_i^{\uparrow}(e^{G_1+G_2})\right)^p.$$

Proof. By the Lie–Trotter formula, $e^{A+B} = \lim_{n\to\infty}(e^{\frac{A}{n}}e^{\frac{B}{n}})^n$. As both the determinant and the singular values depend continuously on the matrix, we can combine this with (an iterative application of) Lemma IV.1 to see whether it suffices to have $|\det(e^{\frac{G_i}{n}})| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(e^{\frac{G_i}{n}})\right)^p$ for arbitrary $n \in \mathbb{N}$. We can summarize this reasoning as follows:

$$\begin{aligned}|\det(e^{G_1+G_2})| &= \lim_{n\to\infty}|\det\left(\left(e^{\frac{G_1}{n}}e^{\frac{G_2}{n}}\right)^n\right)| \\ &\leq \lim_{n\to\infty}\left(\prod_{i=1}^{k} s_i^{\uparrow}\left(\left(e^{\frac{G_1}{n}}e^{\frac{G_2}{n}}\right)^n\right)\right)^p \\ &= \left(\prod_{i=1}^{k} s_i^{\uparrow}(e^{G_1+G_2})\right)^p,\end{aligned}$$

where the inequality follows by combining the assumption with Lemma IV.1. □

Remark IV.3. If G_j in Lemma IV.2 are normal matrices, then it is easy to see that the assumed inequality for $n = 1$ already implies the corresponding inequality for any $n \in \mathbb{N}$. In general, however, this implication is not true. This can be seen as considering L and $\frac{1}{2}L$, with L given in Example IV.12. Therefore, we make the assumption for all $n \in \mathbb{N}$. This is also why we formulate Definition III.2 for convex sets of generators that contain the zero-matrix.

Next, we discuss how to reduce an inequality of the form (2) for a single matrix exponential to an inequality of eigenvalues of the exponent.

Lemma IV.4. Suppose that $G \in \mathcal{M}_d$ *satisfies* $\mathrm{Tr}[G+G^*] - p\sum_{i=1}^{k}\lambda_i^{\uparrow}(G+G^*) \leq 0$, *then*

$$|\det(e^G)| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(e^G)\right)^p.$$

Proof. We observe that

$$\prod_{i=1}^{k} s_i^{\uparrow}(e^G) = \frac{|\det(e^G)|}{\prod_{i=1}^{d-k} s_i^{\downarrow}(e^G)} \geq \frac{|\det(e^G)|}{\prod_{i=1}^{d-k} s_i^{\downarrow}(e^{12(G+G^*)})} = \frac{\det(e^{\frac{1}{2}(G+G^*)})}{\prod_{i=1}^{d-k} e^{12\lambda_i^{\downarrow}(G+G^*)}} = \prod_{i=1}^{k} e^{\frac{1}{2}\lambda_i^{\uparrow}(G+G^*)},$$

where we used $\prod_{i=1}^{d-k} s_i^{\downarrow}(e^G) \leq \prod_{i=1}^{d-k} s_i^{\downarrow}(e^{\Re(G)})$ (see p. 259 of Ref. 18) as well as $|\det(e^G)| = \det(e^{\frac{1}{2}(G+G^*)})$, which can be seen via Lie–Trotter. With this, we now obtain

$$|\det(e^G)|^2 = e^{\mathrm{Tr}[G+G^*]} \leq \left(e^{\sum_{i=1}^{k}\lambda_i^{\uparrow}(G+G^*)}\right)^p = \left(\prod_{i=1}^{k} e^{\frac{1}{2}\lambda_i^{\uparrow}(G+G^*)}\right)^{2p} \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(e^G)\right)^{2p},$$

where the first inequality is exactly our assumption. Now we take the square root and obtain the claimed inequality. □

We summarize the results of the foregoing discussion for Markovian divisibility in the following.

Theorem IV.5. *Let* $\mathcal{G} \subseteq \mathcal{M}_d$ *be a set of generators. Let* $T \in \overline{\mathcal{D}}_{\mathcal{G}}$ *and suppose that every* $G \in \mathcal{G}$ *satisfies* $\mathrm{Tr}[G+G^*] - p\sum_{i=1}^{k}\lambda_i^{\uparrow}(G+G^*) \leq 0$. *Then,* $|\det(T)| \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T)\right)^p$.

Proof. By continuity of the determinant and the singular values, we can restrict our attention to $T \in \mathcal{D}_G$. In that case, there exist $n \in \mathbb{N}$ and generators $\{G_i\}_{1\leq i\leq n} \subset \mathcal{G}$ such that $\prod_{i=1}^{n} e^{G_i} = T$. By Lemma IV.1, it suffices to have the desired inequality for each factor e^{G_i}. These now satisfy the inequality by Lemma IV.4.

We obtain an analogous result for infinitesimal Markovian divisibility:

Corollary IV.6. Let $\mathcal{G} \subset \mathcal{M}_d$ *be a compact and convex set of matrices containing* $0 \in \mathcal{M}_d$. *Let* $\tilde{\mathcal{G}} := \{\lambda G \mid \lambda \in [0,1], G \text{ an extreme point of } \mathcal{G}\} \subset \mathcal{G}$. *Assume that every* $\tilde{G} \in \tilde{\mathcal{G}}$ *satisfies* $\mathrm{Tr}[\tilde{G}+\tilde{G}^*] - p\sum_{i=1}^{k}\lambda_i^{\uparrow}(\tilde{G}+\tilde{G}^*) \leq 0$. *Let* $T \in \overline{\mathcal{I}}_{\mathcal{G}}$. *Then,* $0 \leq \det(T) \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T)\right)^p$.

Proof. $\det(T) \geq 0$ follows in the same way as in Proposition III.5. By continuity, it suffices to prove the desired upper bound for $T \in \mathcal{I}_G$. By the definition of the set $\mathcal{I}_G$ and Lemma IV.1, it then suffices to consider single factors of the form e^G, $G \in \mathcal{G}$. By definition of $\tilde{\mathcal{G}}$, $\tilde{G} \in \tilde{\mathcal{G}}$, in particular, implies that $\frac{1}{n}\tilde{G} \in \tilde{\mathcal{G}}$ for all $n \in \mathbb{N}$. In addition, every element of $\mathcal{G}$ can be expressed as a finite sum of elements of $\tilde{\mathcal{G}}$ (by Krein–Milman). Therefore, we can apply Lemma IV.2 to conclude that it suffices to consider single factors of the form $e^{\tilde{G}}$, $\tilde{G} \in \tilde{\mathcal{G}}$. Now we apply Lemma IV.4 to finish the proof. □

The assumption in Corollary IV.6 is about (truncated) rays through extreme points of the convex set of interest. In light of Remark IV.3, we expect that this can, in general, not be further simplified to an assumption only about the extreme points themselves (without multiples).

B. Quantum channels

We now want to apply the reasoning from Subsection IV A to the more specific question of infinitesimal (Markovian) divisibility of quantum channels.

To avoid confusion about notation, in this subsection, we will denote the eigenvalues of a matrix $\mathcal{L}$ as $\lambda_i = \lambda_i(\mathcal{L})$, whereas the eigenvalues of a linear map L on matrices are written as $\Lambda_K = \Lambda_K(L)$. For real eigenvalues of such linear superoperators, we use $\Lambda_K^{\downarrow}$ ($\Lambda_K^{\uparrow}$) to denote the eigenvalues in decreasing (increasing) order.

1. Determinant vs power of the smallest singular value

We first show that purely dissipative Lindblad generators with one Lindbladian satisfy an inequality, as assumed in Lemma IV.4 with only one summand:

Lemma IV.7. *Let $L : \mathcal{M}_d \to \mathcal{M}_d$ and $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$ be a purely dissipative Lindblad generator with one Lindbladian $\mathcal{L} \in \mathcal{M}_d$. Then,*

$$\mathrm{Tr}[L + L^*] - \frac{d}{2}\Lambda_1^{\uparrow}(L + L^*) \leq 0. \tag{4}$$

Proof. We adopt the following convention for vectorization of matrices: If A is an $n \times n$-matrix with column vectors a_i, then $vec(A) = (a_1^T, \ldots, a_n^T)^T$ is the column vector obtained by stacking the columns of A on top of one another. When using $vec(ABC) = (C^T \otimes A)vec(B)$ to rewrite $L + L^*$ as a $d^2 \times d^2$-matrix, we obtain

$$vec(L + L^*) = \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L} - \overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d.$$

From this, it is easy to see that

$$\mathrm{Tr}[L + L^*] = |\mathrm{Tr}[\mathcal{L}]|^2 - 2d\|\mathcal{L}\|_F^2.$$

We observe that the Lindbladians $\mathcal{L}$ and $\lambda\mathbb{1}_d + \mathcal{L}$ give rise to the same superoperator $L + L^*$ for every $\lambda \in \mathbb{C}$. Hence, we can, without loss of generality, assume that $\mathrm{Tr}[\mathcal{L}] = 0$ and therefore $\mathrm{Tr}[L + L^*] = -2d\|\mathcal{L}\|_F^2$. Thus, we obtain

$$\begin{aligned}
\mathrm{Tr}[L + L^*] - \frac{d}{2}\Lambda_1^{\uparrow}(L + L^*) &\leq -2d\|\mathcal{L}\|_F^2 + \frac{d}{2}\|L + L^*\|_\infty \\
&\leq -2d\|\mathcal{L}\|_F^2 + \frac{d}{2}\left(\|\overline{\mathcal{L}} \otimes \mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger\|_\infty + \|\mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d\|_\infty\right) \\
&= -2d\|\mathcal{L}\|_F^2 + \frac{d}{2} \cdot 4\|\mathcal{L}\|_\infty^2 \\
&\leq 0,
\end{aligned}$$

which finishes the proof. □

Remark IV.8. In our Proof of Lemma IV.7, one step might strike the reader as particularly simplistic. Specifically, we estimate

$$\frac{d}{2}\|L + L^*\|_\infty \leq \frac{d}{2}\left(\|\overline{\mathcal{L}} \otimes \mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger\|_\infty + \|\mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L}\|_\infty + \|\overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d\|_\infty\right) \leq \frac{d}{2} \cdot 4\|\mathcal{L}\|_\infty^2.$$

With a more thorough analysis, we can slightly improve this upper bound and thereby increase the prefactor in the statement of Lemma IV.7 from $\frac{d}{2}$ to $\approx 0.610\,733\, d$. (We then get the same improvement in Corollary IV.9.) We derive this improvement in the Appendix.

We can now apply the reasoning from Subsection IV A (for $k = 1$ and $p = \frac{d}{2}$) to obtain the following corollary:

Corollary IV.9. *Let $T \in \overline{\mathcal{I}}_d$. Then, $0 \leq \det(T) \leq \left(s_1^{\uparrow}(T)\right)^{\frac{d}{2}}$.*

Proof. By combining the form of Lindblad generators from Theorem II.2 with Corollary IV.6, it suffices to consider Lindblad generators with a single summand, i.e., of the form

$$L(\rho) = \begin{cases} i[\rho, H] \text{ with } H = H^\dagger \\ \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}. \end{cases}$$

$[\cdot, H] : \mathcal{M}_d \to \mathcal{M}_d$ is a self-adjoint map if $H = H^\dagger$, and therefore, $e^{i[\cdot,H]}$ has 1 as only singular value. The desired singular value inequality (2) is thus trivially satisfied for factors of this form. For factors of the form e^L with $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$, the desired eigenvalue inequality is exactly shown in Lemma IV.7. □

This necessary criterion can be used to find channels that are not infinitesimal divisible and are given by convex combinations of a rank-deficient channel with the identity channel.

Corollary IV.10. Let $T : \mathcal{M}_d \to \mathcal{M}_d$ be a quantum channel that has singular value of 0 of multiplicity $1 \le k < \frac{d}{2}$. Then, every neighborhood of T contains a non-infinitesimal divisible channel.

Proof. Given such a quantum channel T, we can explicitly write down non-infinitesimal divisible channels via convex combination with the identity, $T_\epsilon = (1-\epsilon)T + \epsilon\,\mathrm{Id}$. By assumption, T_ϵ has exactly k singular values, which go to 0 as $\epsilon \to 0$. Thus, either $\det(T_\epsilon) < 0$ or we have

$$\det(T_\epsilon) = \prod_{j=1}^{d^2} s_j^\uparrow(T_\epsilon) \ge \left(s_1^\uparrow(T_\epsilon)\right)^k \prod_{j=k+1}^{d^2} s_j^\uparrow(T_\epsilon) > \left(s_1^\uparrow(T_\epsilon)\right)^{d/2} \quad \text{for } \epsilon \text{ small enough,}$$

where we just used that the $d^2 - k$ largest singular values do not go to 0 for $\epsilon \to 0$. Hence, for $\epsilon > 0$ small enough, T_ϵ does not satisfy the criterion given in Corollary IV.9 and is therefore not infinitesimal divisible. □

Example IV.11. We can use the above Corollary to find infinitesimal divisible channels near the channel $T : \mathcal{M}_d \to \mathcal{M}_d$, $T(\rho) = \frac{\mathrm{Tr}[\rho]}{d}\mathbb{1}_d$. T is diagonal with respect to the generalized Gell–Mann basis of $\mathcal{M}_d$ with the corresponding matrix given by $\hat{T} = \mathrm{diag}[1, 0, 0, \ldots, 0]$. The Choi matrix τ of T has full rank and is thus particularly strictly positive definite (because complete positivity of T translates to positive semidefiniteness of its Choi matrix τ; see Ref. 15).

Hence, we can pick $\varepsilon > 0$ small enough such that $\hat{T}_\varepsilon = \mathrm{diag}[1, \varepsilon, \ldots, \varepsilon, 0]$ is the matrix representation of a completely positive map in the generalized Gell–Mann basis. As such a matrix $\hat{T}_\varepsilon$ describes by its very form a trace-preserving map, it corresponds to a quantum channel T_ε, which now has an eigenvalue of 0 with a multiplicity of 1. Hence, we can apply Corollary IV.10 to T_ε and thus find channels arbitrarily close to T that are not infinitesimal divisible.

Naturally, the question arises whether the power $\frac{d}{2}$ in Corollary IV.9 is optimal. Our next example shows that the dependence on d cannot be better than linear and that the factor of $\frac{1}{2}$ cannot be improved by much.

Example IV.12. When considering the pathological case of a matrix of the form

$$\mathcal{L} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

we can easily compute that

$$L + L^* = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_d, \end{pmatrix}$$

with $D_i = \mathrm{diag}(0,\dots,0,-1) \in \mathbb{R}^{d\times d}$ for $1 \le i \le d-1$ and $D_d = \mathrm{diag}(-1,\dots,-1,-2) \in \mathbb{R}^{d\times d}$. Hence, $L + L^*$ has eigenvalues -1 of multiplicity $2(d-1)$, 0 of multiplicity $d^2 - 2d$, and $-1 \pm \sqrt{2}$, each of multiplicity 1. In particular, $\mathrm{Tr}[L + L^*] - p\Lambda_1^{\uparrow}(L + L^*) = -2d + (1+\sqrt{2})p \le 0$ iff $p \le \frac{2}{1+\sqrt{2}} d$.

This example also shows that in Theorem IV.9, nothing better than $\det(T) \le \left(s_1^{\uparrow}(T)\right)^p$ with $p = \mathcal{O}(d)$ can be achieved. Specifically, with the above choice of $\mathcal{L}$, we get

$$L = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_d \end{pmatrix}.$$

This can now be exponentiated to obtain

$$T := e^L = \begin{pmatrix} 0 & 0 & \cdots & 1-e^{-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} e^{\frac{1}{2}D_1} & 0 & \cdots & 0 \\ 0 & e^{\frac{1}{2}D_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\frac{1}{2}D_d} \end{pmatrix},$$

where $e^{1/2D_i} = \mathrm{diag}(1,\dots,1,e^{-1/2})$ for $1 \le i \le d-1$ and $e^{1/2D_d} = \mathrm{diag}(e^{-1/2},\dots,e^{-1/2},e^{-1})$.

We can now compute

$$T^*T = \begin{pmatrix} 0 & 0 & \cdots & 1-e^{-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 1-e^{-1} & 0 & \cdots & (1-e^{-1})^2 \end{pmatrix} + \begin{pmatrix} e^{D_1} & 0 & \cdots & 0 \\ 0 & e^{D_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{D_d} \end{pmatrix},$$

from which we see that T has singular values of 1 of multiplicity $(d-1)^2 - 1$, $e^{-\frac{1}{2}}$ of multiplicity $2(d-1)$, $\frac{\sqrt{1-e+e^2+(e-1)\sqrt{1+e^2}}}{e} \approx 1.200$ of multiplicity 1, and $\frac{\sqrt{1-e+e^2-(e-1)\sqrt{1+e^2}}}{e} \approx 0.306$ of multiplicity 1. In particular, we have

$$\det(T) \le \left(s_1^{\uparrow}(T)\right)^{\frac{d}{2}},$$

but

$$\det(T) > \left(s_1^{\uparrow}(T)\right)^{d}.$$

More precisely, we see that $\det(T) \le \left(s_1^{\uparrow}(T)\right)^p$ requires, as $d \to \infty$,

$$p \le \frac{\ln(s_1^{\downarrow}(T)) + \ln(s_1^{\uparrow}(T)) - (d-1)}{\ln(s_1^{\uparrow}(T))} \approx \frac{\ln(1.200) + \ln(0.306) - (d-1)}{\ln(0.306)} \sim \frac{1}{-\ln(0.306)} d \approx 0.845\ d.$$

If we do the same computation for $\frac{1}{n}L$ instead of L, we obtain, in the limit of large n, the upper bound,

$$p \le \frac{2}{1+\sqrt{2}}\ d + 1 + \frac{\sqrt{2}}{1+\sqrt{2}},$$

which coincides up to an additive constant with the bound obtained above on the level of eigenvalues.

This concludes our discussion of the example.

The result of Theorem IV.9 applied to the qubit case does not reproduce the criterion from Theorem III.6. In particular, we do not obtain $s_{\min}^2$ but merely $s_{\min}$. For normal Lindbladians and thus products of unital channels, our reasoning can, however, be improved.

Proposition IV.13. For normal Lindbladians, the prefactor in Lemma IV.7 (thus the exponent in Corollary IV.9) can be improved to d. Furthermore, this estimate is sharp, i.e., *cannot be improved for general normal $\mathcal{L}$.*

Proof. For normal $\mathcal{L}$, we know all the eigenvalues of $L + L^*$, and they are given by $\{-|\lambda_i - \lambda_j|^2\}_{i,j}$, where λ_i are the eigenvalues of $\mathcal{L}$ (see Remark IV.15 for a detailed derivation). Now choose two indices i^*, j^* such that

$$|\lambda_{i^*} - \lambda_{j^*}|^2 = \max_{i,j} |\lambda_i - \lambda_j|^2.$$

Then, (4) for exponent d becomes

$$\mathrm{Tr}[L + L^*] - d\Lambda_1^{\uparrow}(L + L^*) = -\sum_{i,j} |\lambda_i - \lambda_j|^2 + d|\lambda_{i^*} - \lambda_{j^*}|^2. \tag{5}$$

Now using $|a + b|^2 \le 2\big(|a|^2 + |b|^2\big)$ and denoting the indices $\{1, \ldots, d\} \backslash \{i^*, j^*\} = \{n_1, \ldots, n_{d-2}\}$, we obtain

$$(5) \le -\sum_{i,j} |\lambda_i - \lambda_j|^2 + 2|\lambda_{i^*} - \lambda_{j^*}|^2 + 2\sum_{k=1}^{d-2} \big(|\lambda_i^* - \lambda_{n_k}|^2 + |\lambda_j^* - \lambda_{n_k}|^2\big) \le 0.$$

In the last step, we used that every difference $|\lambda_{i^*/j^*} - \lambda_{n_k}|^2$ appears twice in the first sum.

In order to see that d is also optimal, consider the example $\mathcal{L} = \mathrm{diag}[1, -1, 0, \ldots, 0]$. Here, a straightforward calculation shows that $L + L^*$ has eigenvalues -4 of multiplicity 2, -1 of multiplicity $4(d-2)$, and 0 of multiplicity $2 + (d-2)^2$. Thus,

$$\mathrm{Tr}[L + L^*] = -4d = -d|\lambda_1 - \lambda_2|^2 = d\Lambda_1^{\uparrow}(L + L^*),$$

so d is optimal. □

Note that the example used in the previous proof can also be used to show that for normal $\mathcal{L}$, the exponent in $\det(e^L) \le \big(s_1^{\uparrow}(e^L)\big)^d$ cannot be improved.

2. Determinant vs product of smallest singular values

So far, we have used the ideas from Subsection IV A to derive an upper bound on the determinant of infinitesimal divisible quantum channels in terms of the power of its smallest singular value. Now we focus on the other aspect of Lemma IV.4 and bound the determinant via a product of smallest singular values.

Lemma IV.14. Let $L : \mathcal{M}_d \to \mathcal{M}_d$ and $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$ be a purely dissipative Lindblad generator with one Lindbladian $\mathcal{L} \in \mathcal{M}_d$. Then, for $f(d) = 2d - 2\sqrt{2d} + 1$, we have

$$\mathrm{Tr}[L + L^*] - \sum_{K=1}^{\lfloor f(d) \rfloor} \Lambda_K^{\uparrow}(L + L^*) \le 0. \tag{6}$$

Proof. As in the Proof of Lemma IV.7, we can, without loss of generality, assume that $\mathrm{Tr}[\mathcal{L}] = 0$, and therefore, $\mathrm{Tr}[L + L^*] = -2d\|\mathcal{L}\|_F^2$. We can now bound

$$\begin{aligned}
-\sum_{K=1}^{\lfloor f(d) \rfloor} \Lambda_K^{\uparrow}(L + L^*) &\le \sum_{K=1}^{\lfloor f(d) \rfloor} |\Lambda_K^{\uparrow}(L + L^*)| \\
&\le \sum_{K=1}^{\lfloor f(d) \rfloor} s_K^{\downarrow}(L + L^*) \\
&= \big\|L + L^*\big\|_{(\lfloor f(d) \rfloor)} \\
&= \Big\|\overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L} - \overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d\Big\|_{(\lfloor f(d) \rfloor)} \\
&\le 2\big\|\overline{\mathcal{L}} \otimes \mathcal{L}\big\|_{(\lfloor f(d) \rfloor)} + \Big\|\mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L} + \overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d\Big\|_{(\lfloor f(d) \rfloor)},
\end{aligned}$$

where we used the kth Ky Fan norm,

$$\|A\|_{(k)} := \sum_{i=1}^{k} s_i^{\downarrow}(A).$$

We bound those two norms separately: For the first term,

$$\begin{aligned}\left\|\overline{\mathcal{L}} \otimes \mathcal{L}\right\|_{(\lfloor f(d)\rfloor)} &= \sum_{K=1}^{\lfloor f(d)\rfloor} s_K^{\downarrow}(\overline{\mathcal{L}} \otimes \mathcal{L}) \\ &\leq \sqrt{\lfloor f(d)\rfloor}\left(\sum_{K=1}^{\lfloor f(d)\rfloor}\left(s_K^{\downarrow}(\overline{\mathcal{L}} \otimes \mathcal{L})\right)^2\right)^{\frac{1}{2}} \\ &\leq \sqrt{\lfloor f(d)\rfloor}\left\|\overline{\mathcal{L}} \otimes \mathcal{L}\right\|_F \\ &= \sqrt{\lfloor f(d)\rfloor}\|\mathcal{L}\|_F^2,\end{aligned}$$

where the first inequality is an application of Cauchy–Schwarz.

For the second term, we choose an ONB with respect to which $\mathcal{L}^\dagger\mathcal{L}$ is diagonal with the squares of the singular values s_i of $\mathcal{L}$ on the diagonal (which is possible by unitary invariance of the Ky Fan norms) and then compute

$$\begin{aligned}\left\|\mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L} + \overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d\right\|_{(\lfloor f(d)\rfloor)} &= \left\|\mathrm{diag}[2s_1^2, s_1^2 + s_2^2, \ldots, s_1^2 + s_d^2, s_1^2 + s_2^2, \ldots, 2s_d^2]\right\|_{(\lfloor f(d)\rfloor)} \\ &\leq (\lfloor f(d)\rfloor + 1)\sum_{i=1}^{d} s_i^2 \\ &\leq (\lfloor f(d)\rfloor + 1)\|\mathcal{L}\|_F^2.\end{aligned}$$

Plugging this into the above, we obtain

$$\mathrm{Tr}[L + L^*] - \sum_{K=1}^{\lfloor f(d)\rfloor} \Lambda_K^{\uparrow}(L + L^*) \leq -2d\|\mathcal{L}\|_F^2 + (1 + 2\sqrt{\lfloor f(d)\rfloor} + \lfloor f(d)\rfloor)\|\mathcal{L}\|_F^2.$$

This is ≤ 0 if $1 + 2\sqrt{f(d)} + f(d) - 2d \leq 0$, which is guaranteed by the choice $f(d) = 2d - 2\sqrt{2d} + 1$. □

Remark IV.15. The reasoning in the Proof of Lemma IV.14 becomes particularly simple if the Lindbladian $\mathcal{L}$ is normal. In that case, let $\{v_j\}_j$ be an orthonormal basis for $\mathbb{R}^d$ consisting of eigenvectors of $\mathcal{L}$ corresponding to eigenvalues $\{\lambda_j\}_j$. By normality, the $\{v_j\}_j$ are also eigenvectors of $\mathcal{L}^\dagger$ to eigenvalues $\{\overline{\lambda}_j\}_j$. Recalling that in the matrix representation, we can write $L + L^* = \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger\mathcal{L} - \overline{\mathcal{L}^\dagger\mathcal{L}} \otimes \mathbb{1}_d$, it is now easy to see that $\{\bar{v}_i \otimes v_j\}_{i,j}$ is an orthonormal basis of $\mathbb{C}^{d^2}$ consisting of eigenvectors of $L + L^*$ to eigenvalues $\{-|\lambda_i - \lambda_j|^2\}_{i,j}$. Hence, all eigenvalues of $L + L^*$ are ≤ 0, and the inequality of Lemma IV.14 is trivially satisfied.

We can now apply our reasoning from Subsection IV A (with $k = \lfloor 2d - 2\sqrt{2d} + 1\rfloor$ and $p = 1$) to obtain the following corollary:

Corollary IV.16. Let $T \in \overline{\mathcal{I}}_d$. Then, with $f(d) = \lfloor 2d - 2\sqrt{2d} + 1\rfloor$, we have

$$0 \leq \det(T) \leq \prod_{i=1}^{f(d)} s_i^{\uparrow}(T).$$

Example IV.17. Consider again the Lindblad generator L from Example IV.12 and the corresponding channel T. With the eigenvalues and singular values computed in Example IV.12, we see that in this case, $\sum_{i=1}^{d^2-k} \Lambda_i^{\downarrow}(L + L^*) > 0$ for all $k \geq 2d - 1$, and we have

$$\det(T) \leq \prod_{i=1}^{2d-2} s_i^{\uparrow}(T),$$

but

$$\det(T) > \prod_{i=1}^{k} s_i^{\uparrow}(T)$$

for every $d^2 > k > 2d-2$. This shows that in Corollary IV.16, nothing better than $\det(T) \leq \prod_{i=1}^{k} s_i^{\uparrow}(T)$ with $k = 2d-2$ can be achieved.

Remark IV.18. After establishing the optimality of picking the smallest $2d - C$ singular values in Corollary IV.16, the question naturally arises whether this bound can, in principle, be achieved with our proof strategy. In other words, what is the optimal choice for k such that

$$\left\| \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger \mathcal{L}} \otimes \mathbb{1}_d \right\|_{(k)} \leq 2d \|\mathcal{L}\|_F^2 ?$$

We clearly have

$$\left\| \overline{\mathcal{L}} \otimes \mathcal{L} + \overline{\mathcal{L}^\dagger} \otimes \mathcal{L}^\dagger - \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} - \overline{\mathcal{L}^\dagger \mathcal{L}} \otimes \mathbb{1}_d \right\|_{(k)} \leq 2 \left\| \overline{\mathcal{L}} \otimes \mathcal{L} \right\|_{(k)} + \left\| \mathbb{1}_d \otimes \mathcal{L}^\dagger \mathcal{L} + \overline{\mathcal{L}^\dagger \mathcal{L}} \otimes \mathbb{1}_d \right\|_{(k)}.$$

The first term has the singular values $s_i(\mathcal{L}) s_j(\mathcal{L})$, and the second one has singular values $s_i^2(\mathcal{L}) + s_j^2(\mathcal{L})$. Thus, if we normalize the Frobenius norm of $\mathcal{L}$ to 1 and write $p_i = s_i^2(\mathcal{L})$, we can reduce the desired bound to the following conjecture:

Conjecture IV.19. Let $p \in \mathbb{R}^d_{\geq 0}$ *with* $\sum_{i=1}^{d} p_i = 1$. *Define the matrices* $a, g \in \mathbb{R}^{d \times d}$ *via*

$$a_{ij} = \frac{p_i + p_j}{2}, \quad g_{ij} = \sqrt{p_i p_j}.$$

Denote by $a_k^{\downarrow}$ *and* $g_k^{\downarrow}$ *the* kth *largest entry of* a *and* g, *respectively. Define*

$$A = \sum_{k=1}^{h(d)} a_k^{\downarrow}, \quad G = \sum_{k=1}^{h(d)} g_k^{\downarrow}.$$

We conjecture that the maximal integer $h(d)$ *such that* $A + G \leq d$ *holds for any probability vector* p *is given by* $h(d) = 2d - 5$.

We have tested this conjecture numerically for a wide range of dimensions. Theoretically, it stems from the fact that we know the optimal values and corresponding probability vectors for the arithmetic $[h(d) = 2d-2]$ and geometric mean $[h(d) = d^2]$, respectively. Hence, A is by far more decisive and G can only worsen the maximal number of summands by a bit. If we were able to prove this conjecture, we could choose $f(d) = h(d) = 2d - 5$ in Corollary IV.16, which would bring us closer to the optimum of $2d - 2$ up to an additive constant.

Remark IV.20. In contrast to Subsection IV B 1, here, we cannot provide an example of a quantum channel that violates the criterion from Corollary IV.16. As any channel having only singular values ≤ 1 trivially satisfies the criterion, no unital channel will provide a violation, which makes analytically constructing an example more difficult. We have also tried to find an example of a non-infinitesimal divisible channel that is recognized as such by the conjectured optimal version of our criterion (which we cannot prove yet) numerically via minimizing the fraction $\prod_{i=1}^{2d-2} s_i^{\uparrow}(T) / \det(T)$ over channels. This has, however, not been successful. We would be interested in any comments as to how such an example can be found or why finding one is a challenging task.

So far in our treatment of infinitesimal divisible quantum channels, we considered two extreme cases, namely, estimating the determinant by the highest possible power of the smallest singular value and by the product of the largest possible number of the lowest singular values all with exponent 1. The next proposition corresponds to an interpolation between those two results.

Proposition IV.21. *Let* $T \in \overline{\mathcal{I}}_d$. *Then, for any* $1 \leq k \leq d^2$ *with* $g(d) = \frac{2d}{k + 2\sqrt{k} + 1}$, *we have*

$$0 \leq \det(T) \leq \left(\prod_{i=1}^{k} s_i^{\uparrow}(T) \right)^{g(d)}.$$

Proof. As shown in Subsection IV A, it suffices to show that any Lindblad generator L satisfies

$$\mathrm{Tr}[L + L^*] - g(d) \sum_{\ell=1}^{k} \Lambda_\ell^{\uparrow}(L + L^*) \leq -2d \|\mathcal{L}\|_F^2 + g(d) \left\| L + L^* \right\|_{(k)} \leq 0.$$

Again, we only need to consider purely dissipative Lindblad generators with a single Lindbladian. For such generators, the desired assertion follows from the bound on the Ky Fan norm provided in the Proof of Lemma IV.14,

$$\left\|L + L^*\right\|_{(k)} \leq \left(k + 2\sqrt{k} + 1\right)\|\mathcal{L}\|_F^2.$$

□

Remark IV.22. In our numerical tests, we observe the result of Corollary IV.9 to be the strongest in generic cases in higher dimensions, since generically, the smallest singular value seems to be of some orders of magnitude smaller than the others. However, the result in Proposition IV.21 might give useful improvements for small dimensions, especially if some of the lowest singular values are all of the same order of magnitude. Take the case $d = 3, k = 2$, and then, we get the three results,

$$0 \leq \det(T) \leq \begin{cases} s_1^{\uparrow}(T)^{3/2} & \text{(Corollary IV.9)} \\ s_1^{\uparrow}(T)s_2^{\uparrow}(T) & \text{(Corollary IV.16)} \\ \left(s_1^{\uparrow}(T)s_2^{\uparrow}(T)\right)^{\frac{6}{3+2\sqrt{2}}} & \text{(Proposition IV.21).} \end{cases}$$

Hence, if $s_1^{\uparrow}(T)$ is a lot smaller than $s_2^{\uparrow}(T)$, the first result is the strongest. However, if $s_1^{\uparrow}(T) \approx s_2^{\uparrow}(T)$, then the last result becomes the strongest criterion out of the three.

C. Stochastic matrices

The classical counterparts of quantum channels and Lindblad generators are stochastic matrices and transition rate matrices, respectively. In particular, when choosing the set of generators to be the set of all transition rate matrices, we obtain a notion of (infinitesimal) Markovian divisibility for stochastic matrices.

Motivated by the results of Subsections IV A and IV B, we now study whether similar criteria for infinitesimal divisibility of stochastic matrices can be established. More precisely, we define the following:

Definition IV.23 (Markovian divisible stochastic matrices). We define the set of $d \times d$ stochastic matrices to be

$$\mathcal{S}_d := \left\{ S \in \mathbb{R}^{d\times d} \mid S_{ij} \geq 0 \ \forall i,j \ \text{and} \ \sum_{j=1}^{d} S_{ij} = 1 \ \forall i \right\}$$

and the set of $d \times d$ transition rate matrices to be

$$\mathcal{Q}_d := \left\{ Q \in \mathbb{R}^{d\times d} \mid Q_{ij} \geq 0 \ \forall i \neq j \ \text{and} \ \sum_{j=1}^{d} Q_{ij} = 0 \ \forall i \right\}.$$

We call a stochastic matrix $S \in \mathcal{S}_d$ Markovian divisible *if it is Markovian divisible with respect to the set of generators $\mathcal{Q}_d$ in the sense of Definition III.1.*

Note that, as discussed in Remark III.4, the "infinitesimal" requirement is automatically contained in this definition due to the structure of the set $\mathcal{Q}_d$, which is why we do not write it out explicitly.

Our first observation is that, in contrast to the case of Lindblad generators studied in Subsection IV B, when allowing all transition rate matrices as generators, no non-trivial necessary criteria of our desired form (2) can hold.

Example IV.24. Take the transition rate matrix

$$Q = \begin{pmatrix} -1 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathcal{Q}_d, \quad \text{and then,} \ e^Q = \begin{pmatrix} \frac{1}{e} & 0 & \cdots & 0 & 1-\frac{1}{e} \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

which has singular values $\frac{\sqrt{1-e+e^2+(e-1)\sqrt{1+e^2}}}{e} \approx 1.200$ of multiplicity 1, 1 of multiplicity $d-2$, and $\frac{\sqrt{1-e+e^2-(e-1)\sqrt{1+e^2}}}{e} \approx 0.306$ of multiplicity 1. In particular, we see that for every $1 \leq k < d$,

$$\det(e^Q) > \prod_{i=1}^{k} s_i^{\uparrow}(e^Q).$$

Hence, for Markovian divisible stochastic matrices, there cannot be a non-trivial necessary criterion of the form of Corollary IV.16. Similarly, no non-trivial necessary criterion as in Corollary IV.9 with an exponent growing with some positive power of d can hold when we take the set $\mathcal{G}$ of generators to be all transition rate matrices.

This example, together with Corollaries IV.16 and IV.9, implies the following:

Corollary IV.25. There cannot be a mapping from $d^2 \times d^2$ stochastic matrices to $\mathcal{T}_d$ that both preserves infinitesimal Markovian divisibility and leaves singular values invariant.

We can, however, restrict our attention to strict subsets of all transition rate matrices and derive analogous criteria there.

Lemma IV.26. Let $c \in (0,1]$. Consider the set of generators

$$\mathcal{G}_c := \{Q \in \mathbb{R}^{d\times d} \mid Q \text{ is a transition rate matrix and } Q_{kk} \leq c \min_{1\leq l\leq d} Q_{ll} \ \forall 1 \leq k \leq d\}.$$

Then, $\mathrm{Tr}[Q+Q^T] - \frac{1+c(d-1)}{2}\lambda_1^{\uparrow}(Q+Q^T) \leq 0$.

Proof. Clearly, for $Q \in \mathcal{G}_c$, we have $\mathrm{Tr}[Q+Q^T] = 2\sum_{i=1}^{d} Q_{ii} \leq 2(1+c(d-1)) \min_{1\leq l\leq d} Q_{ll}$. As $\sum_{j=1}^{d} Q_{ij} = 0$ for all $1 \leq i \leq d$, we can use Gerschgorin discs to obtain $\lambda_1^{\uparrow}(Q+Q^T) \geq 4 \min_{1\leq l\leq d} Q_{ll}$. In particular, we have that

$$\mathrm{Tr}[Q+Q^T] - \frac{1+c(d-1)}{2}\lambda_1^{\uparrow}(Q+Q^T) \leq 2(1+c(d-1)) \min_{1\leq l\leq d} Q_{ll} - 2(1+c(d-1)) \min_{1\leq l\leq d} Q_{ll} = 0,$$

as claimed. □

According to our reasoning from Subsection IV A, this directly implies the following corollary:

Corollary IV.27. Let $c \in (0,1]$. Suppose that $S \in [0,1]^{d\times d}$ is a stochastic matrix that is Markovian divisible with respect to $\mathcal{G}_c$. Then, $\det(S) \leq \left(s_1^{\uparrow}(S)\right)^{\frac{1+c(d-1)}{2}}$.

If we set $c = 1$, then $\mathcal{G}_1$ describes the set of transition rate matrices with constant diagonal. For Markovian divisibility of a stochastic matrix S with respect to this restricted set of generators, we obtain again the criterion $\det(S) \leq \left(s_1^{\uparrow}(S)\right)^{\frac{d}{2}}$.

V. CONCLUSION

In this work, we described how the notion of infinitesimal Markovian divisibility introduced in Ref. 3 as a notion of Markovianity for quantum channels with the generators in Lindblad form can be extended to a notion applicable to general linear maps and a (closed and convex) set of generators.

Our main contribution toward an understanding of this notion is a general proof strategy based on (sub-) multiplicativity properties of the determinant and products of largest singular values as well as Trotterization, with which we can establish necessary criteria for infinitesimal Markovian divsibility from a spectral property of the generators.

We showed that all Lindblad generators satisfy such a property, and therefore, our approach yields necessary criteria for infinitesimal Markovian divisibility of quantum channels in any (finite) dimension. These are the first such criteria beyond dimension 2 aside from non-negativity of the determinant. Using these criteria, we gave new examples of provably non-infinitesimal Markovian divisible quantum channels that can be found in any neighborhood of any rank-deficient quantum channel.

However, when studying the classical counterpart—stochastic matrices as maps of interest and transition rate matrices as generators—we found that in the general scenario in which all possible transition rate matrices are allowed as generators, no necessary criterion of our desired form can hold. We could apply our proof strategy only after imposing an additional restriction on the allowed transition rate matrices, which

can be interpreted as requiring that the time scales for remaining in any of the states of the Markov chain are comparable. (In particular, we have to assume that there are no absorbing states.)

Several follow-up questions arise naturally from our work. The first such question is for improvements of our results of Corollaries IV.9 and IV.16. In Examples IV.12 and IV.17, we have shown that our results are close to optimal with respect to the dimension dependence of the exponent in Corollary IV.9 and optimal in the leading order with respect to the number of factors in Corollary IV.16. Nevertheless, there remains a gap to be closed. One possible step for improving Corollary IV.16 might lie in a better understanding of Conjecture IV.19. One might also wonder whether there is a subclass of Lindblad operators for which our proof strategy yields stronger bounds.

More generally, we are hoping for a better understanding of the result of Corollary IV.16. A crucial first step would be to find—either analytically or numerically—examples of not infinitesimal Markovian divisible quantum channels that violate the inequality in Corollary IV.16 (or, for that matter, our conjectured improvement of it). As our proof of this inequality makes extensive use of the assumed divisibility structure, we would consider it surprising if no such examples could be found, which would make it trivial as a necessary criterion.

We mention one more natural question concerning the case of infinitesimal Markovian divisible quantum channels. Specifically, now that we have established necessary criteria for this property, can these be complemented by sufficient criteria of a similar form? The results of Ref. 3 show that for generic qubit channels, an inequality between the determinant of a channel and the square of its smallest singular value is indeed a necessary and sufficient criterion for infinitesimal Markovian divisibility. However, it is not at all clear whether this gerneralizes to higher dimensions.

Finally, here, we have applied our general proof strategy to two scenarios: that of Lindblad generators and that of transition rate matrices as generators. It would be interesting to find other sets of matrix semigroups whose generators satisfy a spectral property as required in Theorem IV.5.

ACKNOWLEDGMENTS

M.C.C. and B.R.G. thank Michael M. Wolf for suggesting this problem and for many insightful discussions. We are also grateful for the suggestions made by the anonymous reviewer at the Journal of Mathematical Physics.

M.C.C. gratefully acknowledges support from the TopMath Graduate Center of the TUM Graduate School at the Technical University of Munich, Germany, and the TopMath Program at the Elite Network of Bavaria. M.C.C. was supported by a doctoral scholarship of the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

B.R.G. gratefully acknowledges support from the International Research Training Group (IGDK Munich—Graz) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project No. 188264188/GRK1754.

APPENDIX: PROOF OF AN IMPROVEMENT TO COROLLARY IV.9

As mentioned in Remark IV.8, we are able to improve the exponent in Corollary IV.9 from $\frac{d}{2}$ to $\frac{2}{2+\sqrt{\frac{13}{8}}}\ d \approx 0.610\,733\ d$.

The idea behind the improvement is to estimate more carefully the smallest ("most negative") eigenvalue $\Lambda_1^{\uparrow}(L+L^*)$. In the proof of Corollary IV.9, we simply estimate $\Lambda_1^{\uparrow}(L+L^*)$ from below by $-4\|\mathcal{L}\|_F^2$, which yields the exponent $\frac{d}{2}$ when comparing it to the $-2d\|L\|_F^2$ from the trace of $L+L^*$. To obtain our improved version, we prove the following lemma.

Lemma A.1. Let $\mathcal{L} \in \mathcal{M}_d$ and $L(\rho) = \mathcal{L}\rho\mathcal{L}^\dagger - \frac{1}{2}\{\mathcal{L}^\dagger\mathcal{L}, \rho\}$. Then,

$$\Lambda_1^{\uparrow}(L+L^*) \geq -\left(2+\sqrt{\frac{13}{8}}\right)\|\mathcal{L}\|_F^2.$$

Proof. The starting point for our reasoning is the l^2-version of the Gerschgorin disc Theorem (see Ref. 18), which states that for a Hermitian matrix $A = (a_{ij})_{i,j}$, each interval $[a_{ii} - r_i, a_{ii} + r_i]$ contains at least one eigenvalue of A, where

$$r_i = \left(\sum_{j\neq i}|a_{ij}|^2\right)^{1/2}.$$

Next, note that due to the tensor-structure of $L+L^*$, we can write its entries in a matrix representation as

$$\left(L+L^*\right)_{kl} = \overline{\mathcal{L}}_{(q+1)(p+1)}\mathcal{L}_{rs} + \mathcal{L}_{(p+1)(q+1)}\overline{\mathcal{L}}_{sr} - \delta_{qp}(\mathcal{L}^\dagger\mathcal{L})_{rs} - (\overline{\mathcal{L}^\dagger\mathcal{L}})_{(q+1)(p+1)}\delta_{rs},$$

where $k = qd + r, l = pd + s$ with $q \in \{0, \ldots, d-1\}, r \in \{1, \ldots, d\}$. If we now choose an orthonormal basis such that $\mathcal{L}^\dagger\mathcal{L} = \mathrm{diag}[\sigma_1^2, \ldots, \sigma_d^2]$, we obtain, for the diagonal entries,

$$\left(L+L^*\right)_{kk} = \overline{\mathcal{L}}_{(q+1)(q+1)}\mathcal{L}_{rr} + \mathcal{L}_{(q+1)(q+1)}\overline{\mathcal{L}}_{rr} - \sigma_r^2 - \sigma_{(q+1)}^2.$$

For the off-diagonal entries, we need to consider only the first two terms in $L + L^*$ due to the choice of our basis, i.e., we get, for $k \neq l$,

$$\left(L+L^*\right)_{kl} = \overline{\mathcal{L}}_{(q+1)(p+1)}\mathcal{L}_{rs} + \mathcal{L}_{(p+1)(q+1)}\overline{\mathcal{L}}_{sr}.$$

We need to distinguish two cases.

Case $k = 1$: Here, we have

$$\left(L+L^*\right)_{11} = 2|\mathcal{L}_{11}|^2 - 2\sigma_1^2$$

and

$$\begin{aligned}\sum_{k\neq 1}\left|\left(L+L^*\right)_{1k}\right|^2 &= \sum_{q,r}\left|\overline{\mathcal{L}}_{1(q+1)}\mathcal{L}_{1r} + \mathcal{L}_{(q+1)1}\overline{\mathcal{L}}_{r1}\right|^2 \\ &\leq \sum_q\left|\overline{\mathcal{L}}_{1(q+1)}\right|^2\sum_r\left|\overline{\mathcal{L}}_{1r}\right|^2 + \sum_q\left|\overline{\mathcal{L}}_{(q+1)1}\right|^2\sum_r\left|\overline{\mathcal{L}}_{r1}\right|^2 + 2\left(\sum_r \underbrace{\left|\mathcal{L}_{1r}\overline{\mathcal{L}_{r1}}\right|}_{\leq\frac{1}{2}\left(|\mathcal{L}_{1r}|^2+|\mathcal{L}_{r1}|^2\right)}\right)^2 \\ &\leq \|\mathcal{L}\|_F^2\left(\|\mathcal{L}\|_F^2 + |\mathcal{L}_{11}|^2\right) + \frac{1}{2}\left(\|\mathcal{L}\|_F^2 + |\mathcal{L}_{11}|^2\right)^2,\end{aligned}$$

where in the last step, we used that, since we are summing up the first row and column, only the diagonal entry $|\mathcal{L}_{11}|^2$ appears twice and the sum of the remaining squares can be bounded by one Frobenius norm.

Before we proceed, let us note that without loss of generality, we can normalize $\|\mathcal{L}\|_F^2 = 1$ to make the following computations more readable. Then, we obtain, by completing the square,

$$\sum_{k\neq 1}\left|\left(L+L^*\right)_{1k}\right|^2 \leq 1 + |\mathcal{L}_{11}|^2 + \frac{1}{2}\left(1+|\mathcal{L}_{11}|^2\right)^2 = \left(\sqrt{\frac{3}{2}} + \sqrt{\frac{2}{3}}|\mathcal{L}_{11}|^2\right)^2 - \frac{1}{6}|\mathcal{L}_{11}|^4.$$

Thus,

$$\left(L+L^*\right)_{11} - \left(\sum_{k\neq 1}\left|\left(L+L^*\right)_{1k}\right|^2\right)^{1/2} \geq 2|\mathcal{L}_{11}|^2 - 2\sigma_1^2 - \sqrt{\frac{3}{2}} - \sqrt{\frac{2}{3}}|\mathcal{L}_{11}|^2 \geq -\left(2+\sqrt{\frac{3}{2}}\right).$$

Hence, in this case, we are even able to bound $a_{ii} - r_i$ from below by $-(2+\sqrt{\frac{3}{2}})\|\mathcal{L}\|_F^2$.

Case $k \neq 1$: Here, we obtain, for the diagonal entries using Young's inequality,

$$\left(L+L^*\right)_{kk} = \overline{\mathcal{L}}_{(q+1)(q+1)}\mathcal{L}_{rr} + \mathcal{L}_{(q+1)(q+1)}\overline{\mathcal{L}}_{rr} - \sigma_r^2 - \sigma_{(q+1)}^2 \geq -2\left|\overline{\mathcal{L}}_{(q+1)(q+1)}\mathcal{L}_{rr}\right| - \|\mathcal{L}\|_F^2.$$

Note that the two singular values might be the same but can, nevertheless, be bounded by just one Frobenius norm, which is the important difference to the case $k = 1$.

For the off-diagonal entries, we start off in the same way as above,

$$\begin{aligned}\sum_{l\neq k}\left|\left(L+L^*\right)_{kl}\right|^2 &\leq \sum_{(p,s)\neq(q,r)}\left|\overline{\mathcal{L}}_{(q+1)(p+1)}\mathcal{L}_{rs}\right|^2 + \left|\mathcal{L}_{(p+1)(q+1)}\overline{\mathcal{L}}_{sr}\right|^2 + 2\left|\overline{\mathcal{L}}_{(q+1)(p+1)}\mathcal{L}_{rs}\mathcal{L}_{(p+1)(q+1)}\overline{\mathcal{L}}_{sr}\right| \\ &= \left(\sum_p|\mathcal{L}_{(q+1)(p+1)}|^2\right)\left(\sum_s|\mathcal{L}_{rs}|^2\right) + \left(\sum_p|\mathcal{L}_{(p+1)(q+1)}|^2\right)\left(\sum_s|\mathcal{L}_{sr}|^2\right) \\ &\quad + 2\left(\sum_p\left|\overline{\mathcal{L}}_{(q+1)(p+1)}\mathcal{L}_{(p+1)(q+1)}\right|\right)\left(\sum_s|\mathcal{L}_{rs}\overline{\mathcal{L}}_{sr}|\right) - 4|\overline{\mathcal{L}}_{(q+1)(q+1)}\mathcal{L}_{rr}|^2 \\ &\leq \|\mathcal{L}\|_F^2\left(\|L\|_F^2 + \min\{|\mathcal{L}_{rr}|^2, |\mathcal{L}_{(q+1)(q+1)}|^2\}\right) - 4|\overline{\mathcal{L}}_{(q+1)(q+1)}\mathcal{L}_{rr}|^2 \\ &\quad + \frac{1}{2}\left(\|L\|_F^2 + |\mathcal{L}_{rr}|^2\right)\left(\|L\|_F^2 + |\mathcal{L}_{(q+1)(q+1)}|^2\right).\end{aligned}$$

Again normalizing $\|\mathcal{L}\|_F^2 = 1$ and denoting $x = |\mathcal{L}_{(q+1)(q+1)}|, y = |\mathcal{L}_{rr}|$ give us

$$\left(L + L^*\right)_{kk} - \left(\sum_{l\neq k} \left|\left(L + L^*\right)_{kl}\right|^2\right)^{1/2} \geq -2xy - 1 - \left(\left(1 + \min\{x^2, y^2\}\right) + \frac{1}{2}(1 + x^2)(1 + y^2) - 4x^2y^2\right)^{1/2} =: g(x, y).$$

Taking the minimum of the function on the right-hand side over (the upper half of) the unit disk $x^2 + y^2 \leq 1$ gives us

$$\left(L + L^*\right)_{kk} - \left(\sum_{l\neq k} \left|\left(L + L^*\right)_{kl}\right|^2\right)^{1/2} \geq \min_{B_1(0)} g(x, y) = g\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = -2 - \sqrt{\frac{13}{8}}.$$

As the second case $k \neq 1$ gives us the worse bound, our final estimate is precisely the statement from Lemma A.1. □

Again, this has to be compared to $-2d\|\mathcal{L}\|_F^2$ in the reasoning of the proof of Corollary IV.9, whereby we obtain the claimed exponent $\frac{2}{2+\sqrt{\frac{13}{8}}}\, d$ (instead of the previous $\frac{d}{2}$).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan, J. Math. Phys. **17**, 821 (1976).
[2] G. Lindblad, Commun. Math. Phys. **48**, 119 (1976).
[3] M. M. Wolf and J. I. Cirac, Commun. Math. Phys. **279**, 147 (2008).
[4] T. S. Cubitt, J. Eisert, and M. M. Wolf, Commun. Math. Phys. **310**, 383 (2012).
[5] J. Bausch and T. Cubitt, Linear Algebra Appl. **504**, 64 (2016).
[6] M. M. Wolf, J. Eisert, T. S. Cubitt, and J. I. Cirac, Phys. Rev. Lett. **101**, 150402 (2008).
[7] Á. Rivas, S. F. Huelga, and M. B. Plenio, Rep. Prog. Phys. **77**, 094001 (2014).
[8] H.-P. Breuer, E.-M. Laine, J. Piilo, and B. Vacchini, Rev. Mod. Phys. **88**, 021002 (2016).
[9] L. Li, M. J. W. Hall, and H. M. Wiseman, Phys. Rep. **759**, 1 (2018).
[10] C.-F. Li, G.-C. Guo, and J. Piilo, Europhys. Lett. **127**, 50001 (2019).
[11] D. Davalos, M. Ziman, and C. Pineda, Quantum **3**, 144 (2019).
[12] D. Chruscinski and U. Chakraborty, New J. Phys. **23**, 013009 (2021).
[13] A. Rivas, S. F. Huelga, and M. B. Plenio, Phys. Rev. Lett. **105**, 050403 (2010).
[14] H.-P. Breuer, E.-M. Laine, and J. Piilo, Phys. Rev. Lett. **103**, 210401 (2009).
[15] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 10th ed. (Cambridge University Press, Cambridge, 2010).
[16] F. Verstraete and H. Verschelde, "On quantum channels," eprint arXiv:quant-ph/0202124 (2002).
[17] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1991).
[18] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics Vol. 169 (Springer, New York, NY, 1997).

A.3 Binary classification with classical instances and quantum labels

Binary classification with classical instances and quantum labels

Matthias C. Caro

Binary classification is one of the central and most well studied tasks in classical machine learning theory and practice. There, the challenge is to produce, given input data with (classical) labels 0 or 1, a hypothesis that predicts the labels of previously unseen data points well. This task is formalized in the *probably approximately correct (PAC)* model of binary classification. In this work, we propose a quantum version of PAC binary classification, in which the labels are quantum states. We show how to reduce this problem to a task of classical binary classification with noisy labels, which allows us to prove sample complexity upper bounds. With an information-theoretic proof strategy, we also establish almost matching sample complexity lower bounds. This demonstrates that our suggested semiclassical strategy is effectively optimal for this problem from a sample complexity perspective.

We begin the article with an introductory section, in which we give an overview over our results, the proof strategy, as well as related work. Section 2 first recalls basic notions from quantum information theory (Section 2.1), and then presents the PAC framework for classical binary classification, together with the characterization of its sample complexity in terms of the VC-dimension (Section 2.2).

After these two preparatory sections, we introduce the quantum learning problem investigated in this work in Section 3. Here, the task is to use training data to learn a mapping from classical inputs to quantum states, where the performance is measured in terms of the trace distance between output states. Crucially, the training data is assumed to be classical-quantum: A single training example consists of a classical input and a copy of the corresponding quantum label state. This differentiates our learning problem from scenarios in which classical descriptions of the quantum objects in the training data are provided. We do, however, assume that (classical descriptions of) the two possible quantum label states are known in advance.

Section 4 contains our first results, namely sample complexity upper bounds for binary classification with classical instances and quantum labels. First, in Section 4.1, we treat the agnostic case. We describe a semi-classical learning strategy in which we perform local Holevo-Helstrom measurements on the quantum part of the training data and then classically process the obtained measurement outcomes with a learning algorithm that corrects for label noise. By proving a Rademacher complexity upper bound (Lemma 1 and the discussion thereafter), we establish our sample complexity upper bound for this scenario. Next, Subsection 4.2 contains our sample complexity upper bounds for the realizable case. Again, we reduce the task to a classical problem of learning from noisy labels. For the latter, we determine the optimal sample complexity in Appendix 4 and then translate this guarantee to our quantum learning scenario (Theorem 2).

In Section 5, we complement the results of the previous section with sample complexity lower bounds, assuming that the quantum labels are pure states. Again, we treat the agnostic (Corollary 1) and the realizable case (Theorem 4) separately. The proof strategy for both cases is information-theoretic: We identify pathological distributions such that the ability to solve the quantum learning problem with respect to these distributions implies the ability to extract a cer-

tain amount of information from the training data. Exactly this information content, however, can also be upper bounded using the form of the training data state, as we show with detailed computations in Appendix 1. Comparing lower and upper bounds on the information content in the data leads to our sample complexity lower bounds.
We conclude the article in Section 6 by emphasizing again how our scenario differs from prior work and with some open questions.
The idea for this project was motivated by discussions between my doctoral advisor, Michael M. Wolf, and myself. The formulation of the learning problem studied in this work has arisen from these discussions. I have developed the idea for Examples 1, 2 and 3 in a discussion with Benedikt R. Graswald. I am solely responsible for the scientific content of this article, with the two restrictions just mentioned. As the single author of this article, I am solely responsible for writing this article.

Permission to include:

Matthias C. Caro.
Binary classification with classical instances and quantum labels.
Quantum Mach. Intell. 3, 18 (2021). https://doi.org/10.1007/s42484-021-00043-z.

Springer www.springer.com

Permissions

Get permission to reuse Springer Nature content

Springer Nature is partnered with the Copyright Clearance Center to meet our customers' licensing and permissions needs.

Copyright Clearance Center's RightsLink® service makes it faster and easier to secure permission for the reuse of Springer Nature content to be published, for example, in a journal/magazine, book/textbook, coursepack, thesis/dissertation, annual report, newspaper, training materials, presentation/slide kit, promotional material, etc.

Simply visit SpringerLink and locate the desired content;

Go to the article or chapter page you wish to reuse content from. (Note: permissions are granted on the article or chapter level, not on the book or journal level). Scroll to the botton of the page, or locate via the side bar, the "Reprints and Permissions" link at the end of the chapter or article.

Select the way you would like to reuse the content;

Complete the form with details on your intended reuse. Please be as complete and specific as possible ao as not to delay your permission request;

Create an account if you haven't already. A RightsLink account is different than a SpringerLink account, and is necessary to receive a licence regardless of the permission fee. You will receive your licence via the email attached to your RightsLink receipt;

Accept the terms and conditions and you're done!

For questions about using the RightsLink service, please contact Customer Support at Copyright Clearance Center via phone +1-855-239-3415 or +1-978-646-2777 or email springernaturesupport@copyright.com.

How to obtain permission to reuse Springer Nature content not available online on SpringerLink

Requests for permission to reuse content (e.g. figure or table, abstract, text excerpts) from Springer Nature publications currently not available online must be submitted in writing. Please be as detailed and specific as possible about what, where, how much, and why you wish to reuse the content.

Your contacts to obtain permission for the reuse of material from:

- books: bookpermissions@springernature.com
- journals: journalpermissions@springernature.com

Author reuse

Please check the Copyright Transfer Statement (CTS) or Licence to Publish (LTP) that you have signed with Springer Nature to find further information about the reuse of your content.

Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.
Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]'

If you are any doubt about whether your intended re-use is covered, please contact journalpermissions@springernature.com for confirmation.

Self-Archiving

- Journal authors retain the right to self-archive the final accepted version of their manuscript. Please see our self-archiving policy for full details:
https://www.springer.com/gp/open-access/authors-rights/self-archiving-policy/2124

- Book authors please refer to the information on this link:
https://www.springer.com/gp/open-access/publication-policies/self-archiving-policy

Quantum Machine Intelligence (2021) 3: 18
https://doi.org/10.1007/s42484-021-00043-z

RESEARCH ARTICLE

Binary classification with classical instances and quantum labels

Matthias C. Caro[1,2]

Received: 21 December 2020 / Accepted: 5 March 2021 / Published online: 5 May 2021

Abstract

In classical statistical learning theory, one of the most well-studied problems is that of binary classification. The information-theoretic sample complexity of this task is tightly characterized by the Vapnik-Chervonenkis (VC) dimension. A quantum analog of this task, with training data given as a quantum state has also been intensely studied and is now known to have the same sample complexity as its classical counterpart. We propose a novel quantum version of the classical binary classification task by considering maps with classical input and quantum output and corresponding classical-quantum training data. We discuss learning strategies for the agnostic and for the realizable case and study their performance to obtain sample complexity upper bounds. Moreover, we provide sample complexity lower bounds which show that our upper bounds are essentially tight for pure output states. In particular, we see that the sample complexity is the same as in the classical binary classification task w.r.t. its dependence on accuracy, confidence and the VC-dimension.

Keywords Quantum learning theory · Sample complexity · Binary classification · VC-dimension

1 Introduction

The fields of machine learning and of quantum computation provide new ways of looking at computational problems and have seen a significant increase in academic as well as practical interest since their origins in the 1970s and 1980s. More recently, attention was directed to paths for combining ideas from these two fruitful research areas. This gave rise to new approaches under different names such as "quantum machine learning" or "quantum learning theory".

In classical statistical learning theory, one of the most influential frameworks is that of probably approximately correct (PAC) learning due to Vapnik and Chervonenkis (1971) and Valiant (1984). It is particularly well studied for the task of binary classification. For this problem the so-called VC-dimension Vapnik and Chervonenkis (1971) is known to characterize the sample complexity of learning a function class (Blumer et al. 1989; Hanneke 2016). Motivated by these strong theoretical results, a quantum analog of this problem was soon defined and studied in a series of papers (an overview over which is given in Arunachalam and de Wolf (2017)), which culminated in the results of Arunachalam and de Wolf (2018). There it is shown that the information-theoretic complexity of the task of quantum PAC learning a 0-1-valued function class is characterized by the VC-dimension in exactly the same way as for the classical scenario.

The scenario studied in Arunachalam and de Wolf (2018) assumes the training data available to the learner to be given in a specific quantum form and allows the learner to perform quantum computational operations on that training data. The functions to be learned, however, still map classical inputs to classical outputs. We propose a different quantum version of the binary classification task by not only considering the possibility of quantum training data but by allowing the objects to be learned to be inherently quantum. More specifically, we consider functions that map classical inputs to one of two possible quantum output states ("quantum labels"). These maps describe state preparation procedures. A more general learning task of this type, for which our problem can be seen as a toy model, could be relevant for cases in which state preparation is either costly or time-consuming, e.g., preparing thermal states at low temperatures (see Brandão and Kastoryano 2019; Chowdhury 2020, and references therein). Here, one could first produce sample data, learn a predictor, and then reproduce the preparation more efficiently using the predictor.

✉ Matthias C. Caro
caro@ma.tum.de

[1] Department of Mathematics, Technical University of Munich, Garching, Germany

[2] Munich Center for Quantum Science and Technology (MCQST), Munich, Germany

1.1 Main results

We consider maps $f : \mathscr{X} \to \{\sigma_0, \sigma_1\}$ that assign to points in a classical input space $\mathscr{X}$ one of two labelling quantum states $\{\sigma_0, \sigma_1\}$. (Here, σ_0 and σ_1 are, in general, mixed states described by density matrices.) Let $\mathscr{F}$ be a function class consisting of such functions. We assume the training data to be given as a classical-quantum state about which, according to the laws of quantum theory, we can only gain information by performing measurements.

Our learning model is that of PAC-learning with accuracy ε and confidence δ. Here, we require a learning algorithm, given as input classical-quantum training data generated according to some unknown underlying distribution, to output with probability $\geq 1-\delta$ over the choice of training data a hypothesis that achieves accuracy ε. (Accuracy is measured in terms of the trace distance.)

We present a learning strategy that (ε, δ)-PAC learns $\mathscr{F} \subseteq \{f : \mathscr{X} \to \{\sigma_0, \sigma_1\}\}$ in the agnostic scenario from classical-quantum training data of size $\mathscr{O}\left(\frac{d}{\varepsilon^2} + \frac{\log 1/\delta}{\varepsilon^2}\right)$, where d is the VC-dimension of the $\{0, 1\}$-valued function class $\tilde{\mathscr{F}} \subseteq \{\tilde{f} : \mathscr{X} \to \{0, 1\}\}$ induced by $\mathscr{F}$ via $\sigma_i \mapsto i$, $i = 0, 1$. Here, "agnostic" means that there need not be a function in $\mathscr{F}$ that would achieve perfect accuracy. We also show that solving this learning problem requires training data size $\Omega\left(\frac{d}{\varepsilon^2} + \frac{\log 1/\delta}{\varepsilon^2}\right)$, so our strategy is optimal w.r.t. the sample complexity dependence on ε, δ and d.

For the realizable scenario of the quantum learning problem, i.e., under the assumption that perfect accuracy can be achieved using $\mathscr{F}$, we prove a sample complexity upper bound of

$$\mathscr{O}\left(\frac{1}{\varepsilon(1-2\max\{\mathrm{tr}[E_0\sigma_1], \mathrm{tr}[E_1\sigma_0]\})^2}\,(d+\log 1/\delta)\right),$$

where $\{E_0, E_1\}$ is the Holevo-Helstrom measurement for distinguishing σ_0 and σ_1, and a sample complexity lower bound of $\Omega\left(\frac{d}{\varepsilon} + \frac{\log 1/\delta}{\varepsilon}\right)$. Also here, these bounds coincide w.r.t. their dependence on ε, δ and d. The prefactor $(1 - 2\max\{\mathrm{tr}[E_0\sigma_1],\ \mathrm{tr}[E_1\sigma_0]\})^{-2}$ in the upper bound comes from our procedure trying to distinguish σ_0 and σ_1 by measuring single copies. (Note: Even though we formulate this in terms of the Holevo-Helstrom measurement, we could use any other two-outcome POVM $\{\tilde{E}_0, \tilde{E}_1\}$ that satisfies $\max\{\mathrm{tr}[\tilde{E}_0\sigma_1], \mathrm{tr}[\tilde{E}_1\sigma_0]\} < 1/2$.).

In proving the sample complexity upper bound for the realizable scenario, we combine algorithms from Laird (1988) and Hanneke (2016) to show that $\mathscr{O}\left(\frac{1}{\varepsilon(1-2\eta_b)^2}\right.$ $\left.(d + \log 1/\delta)\right)$ classical examples with two-sided classification noise, i.e., in which each label is flipped with probability given by a noise rate, suffice for classical (ε, δ)-PAC learning a function class of VC-dimension d in the realizable scenario if the noise rate is bounded by $\eta_b < 1/2$. This upper bound has, to the best of our knowledge, not been observed before and, when combined with the lower bound from Arunachalam and de Wolf (2018), establishes the optimal sample complexity of this classical noisy learning problem.

As is common in statistical learning theory, our main focus lies on the information-theoretic complexity of the learning problem, i.e., the necessary and sufficient number of quantum examples, whereas we do not discuss the computational complexity. Our proposed strategies are "semi-classical" in the following sense: After initially performing simple tensor product measurements, in which each tensor factor is a two-outcome POVM, the remaining computation is done by a classical learning algorithm. In particular, the procedure does not require (possibly hard to implement) joint measurements and its computational complexity will be determined by the (classical) computational complexity of the classical learner used as a subroutine.

1.2 Overview over the proof strategy

We first sketch how we obtain the sample complexity upper bounds. We propose a simple (semi-classical) procedure that consists of first performing local measurements on the quantum part of the training data examples to obtain classical training data and then applying a classical learning algorithm.

We observe that the learning problem for which the classical learner is applied, can then be viewed as a classical binary classification problem with two-sided classification noise, i.e., in which the labels are flipped with certain error probabilities determined by the outcome probabilities of the performed quantum measurements. Therefore, we have reduced our problem to obtaining sample complexity upper bounds for a classical learning problem with noise.

In the general (so-called *agnostic*) case, we can use known sample complexity bounds formulated in terms of a complexity measure called *Rademacher complexity* to show that classical empirical risk minimization w.r.t. a suitably modified loss function (as suggested in Natarajan et al. 2013) achieves optimal sample complexity for this classical learning problem with noise.

In the *realizable* case, i.e., under the assumption that any non-noisy training data set can be perfectly represented by some hypothesis in our class $\tilde{\mathscr{F}}$, the optimal sample complexity for binary classification with two-sided classification noise has not been established in the literature. We combine ideas from Laird (1988) and Hanneke (2016) to exhibit an algorithm that achieves information-theoretic optimality for this scenario.

To obtain the sample complexity lower bounds, we apply ideas from Arunachalam and de Wolf (2018). Namely, we

observe that for sufficiently small accuracy parameter, any quantum strategy that solves our learning problem indeed has to be able to distinguish between the possible different training data states with high success probability.

In the simple case of distinguishing between two quantum states, arising from two different "hard-to-distinguish" underlying distributions, this probability can be upper bounded in terms of the trace distance of the states. In the more general case of many states, we do not study this success probability directly. Instead, we consider the information contained in the quantum training data about the choice of the underlying distribution, again chosen out of a set of "hard-to-distinguish" distributions.

1.3 Related work

Bshouty and Jackson (1998) introduced a notion of quantum training data for learning problems with classical concepts and used it to learn DNF (Disjunctive Normal Form) formulae w.r.t. the uniform distribution. This was extended to product distributions by Kanade et al. (2019). Using ideas from Fourier-based learning, this type of quantum training data was also studied in the context of fixed-distribution learning of Boolean linear functions (Bernstein and Vazirani 1993; Cross et al. 2015; Ristè et al. 2017; Grilo et al. 2017; Caro 2020), juntas Atıcı and Servedio (2007), and Fourier-sparse functions (Arunachalam et al. 2019a). Arunachalam and de Wolf (2017) and Arunachalam et al. (2019b) study the limitations of these quantum examples. A broad overview over work on quantum learning classical functions is given in Arunachalam and de Wolf (2017).

Also for the model of learning from membership queries, a quantum counterpart can be considered. Servedio and Gortler (2004) showed that the number of required classical queries is at most polynomially larger than the number of required quantum queries. Recently,this polynomial relation was improved upon in Arunachalam et al. (2019a). A more specific scenario, namely that of learning multilinear polynomials more efficiently from quantum membership queries, is studied in Montanaro (2012).

Similarly, also a quantum counterpart of the classical model of statistical query learning can be defined. This was recently studied in Arunachalam et al. (2020).

Another line of research at the intersection of learning theory and quantum information focuses on applying classical learning to concept classes arising from quantum theory, e.g., from states or measurements. This was initiated by Aaronson (2007) and studied further by Cheng et al. (2016) and Aaronson (2018), and Aaronson et al. (2018).

Our learning model is similar to the one studied in Chung and Lin (2018). Also there, the inputs are assumed to be classical and the outputs are quantum states. The crucial difference to our scenario is that we assume that there are only two possible label states and these are known in advance. In Chung and Lin (2018), there can be a continuum of possible label states.

Our additional assumption allows us to study infinite function classes $\mathscr{F}$, whereas the results in Chung and Lin (2018) are for classes of finite size. (We expect that the reasoning of Chung and Lin (2018) can be extended to infinite classes using the so-called "growth function" when restricting to a finite set of possible target states. This might lead to a learning procedure that can be applied in our scenario without prior knowledge of the possible quantum label states.) As a further difference between the approaches, whereas the strategy of Chung and Lin (2018) requires the ability to perform measurements in random orthonormal bases, the measurements in our procedures can be taken to be fixed and of product form and are thus potentially easier to implement.

The classical problems to which our quantum learning problems are reduced are problems of learning from noisy training data. These were first proposed by Angluin and Laird (1988) and Laird (1988) and studied further, e.g., by Aslam and Decatur (1996) and Cesa-Bianchi et al. (1999) and Natarajan et al. (2013).

1.4 Structure of the paper

In Section 2 we recall some notions from learning theory as well as from quantum information and computation. The central learning problem of this contribution is formulated in Section 3. The next section exhibits strategies for solving the task and establishes sample complexity upper bounds. In doing so, we derive a tight upper bound on the sample complexity of classical binary classification with two-sided classification noise (see Appendix 4). The quantum sample complexity upper bounds are complemented by lower bounds in Section 5. We conclude with open questions and the references.

2 Preliminaries

2.1 Basics of quantum information and computation

A finite-dimensional quantum system is described by a *(mixed)* state and mathematically represented by a *density matrix* of some dimension $d \in \mathbb{N}$, i.e., an element of $\mathscr{S}(\mathbb{C}^d) := \{\rho \in \mathbb{C}^{d\times d} \mid \rho \geq 0, \mathrm{tr}[\rho] = 1\}$. Here, $\rho \geq 0$ means that ρ is a self-adjoint and positive semidefinite matrix. The extreme points of the convex set $\mathscr{S}(\mathbb{C}^d)$ are the rank-1 projections, the *pure states*. We employ Dirac notation to denote a unit vector $\psi \in \mathbb{C}^d$ also by $|\psi\rangle \in \mathbb{C}^d$ and the corresponding pure state by $|\psi\rangle\langle\psi|$.

To make an observation about a quantum system, a measurement has to be performed. Measurements are built from the set of *effect operators* $\mathcal{E}(\mathbb{C}^d) := \{E \in \mathbb{C}^{d\times d} \mid 0 \leq E \leq \mathbb{1}_d\}$. For our purposes it suffices to consider a measurement as a collection $\{E_i\}_{i=1}^{\ell}$ of effect operators $E_i \in \mathcal{E}(\mathbb{C}^d)$ s.t. $\sum_{i=1}^{\ell} E_i = \mathbb{1}_d$. (For the more general notion of a POVM see Nielsen and Chuang (2009) or Heinosaari and Ziman (2012).) When performing a measurement $\{E_i\}_{i=1}^{\ell}$ on a state ρ, output i is observed with probability $\mathrm{tr}[E_i\rho]$. A projective measurement is one where the effect operators are rank-1 projections, i.e., there exists an orthonormal basis $\{|i\rangle\}_{i=1}^{d}$ s.t. $E_i = |i\rangle\langle i|$.

When multiple quantum systems with spaces $\mathbb{C}^{d_i}$ are considered, the composite system is described by the tensor product $\bigotimes_{i=1}^{n} \mathbb{C}^{d_i} \simeq \mathbb{C}^{\prod_i d_i}$ and the set of states becomes $\mathcal{S}(\bigotimes_{i=1}^{n} \mathbb{C}^{d_i})$. Given a state $\rho_{AB} \in \mathcal{S}(\mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B})$ of a composite system, we can obtain states of the subsystems as partial traces $\rho_A = \mathrm{tr}_B[\rho_{AB}]$, $\rho_B = \mathrm{tr}_A[\rho_{AB}]$. Here, the partial trace is defined as satisfying the relation $\mathrm{tr}[(E \otimes \mathbb{1}_{d_B})\rho_{AB}] = \mathrm{tr}[E\mathrm{tr}_B[\rho_{AB}]]$ for all $E \in \mathcal{E}(\mathbb{C}^{d_A})$.

The dynamics of a quantum system are usually described by unitary evolution or, more generally, by quantum channels. For our purposes, these dynamics will not have to be discussed explicitly since they can be considered as part of the performed measurement by changing to the so-called Heisenberg picture (see Nielsen and Chuang 2009). We will take this perspective in proving our sample complexity lower bounds because it allows us to restrict our attention to proving limitations of measurements rather than of channels.

We will also make use of some standard entropic quantities which have been generalized from their classical origins Shannon (1948) to the realm of quantum theory. We denote the Shannon entropy of a random variable X with probability mass function p by $H(X) = -\sum_x p(x)\log(p(x))$, the conditional entropy of a random variable Y given X as $H(Y|X) = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)}\right)$ and the mutual information between X and Y as $I(X:Y) = H(X) + H(Y) - H(X,Y)$. Similarly, the von Neumann entropy of a quantum state ρ will be denoted as $S(\rho) = -\mathrm{tr}[\rho\log\rho]$ and the mutual information for a bipartite quantum state ρ_{AB} as $I(\rho_{AB}) = I(A:B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB})$. All the standard results and inequalities connected to these quantities which appear in our arguments can be found in Nielsen and Chuang (2009) or in Wilde (2013).

2.2 Basics of the PAC framework and the binary classification problem

The setting of *Probably Approximately Correct (PAC)* learning was introduced by Vapnik and Chervonenkis (1971) and Valiant (1984). The general setting is as follows: Let $\mathcal{X}, \mathcal{Y}$ be input and output space, respectively, let $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ be a class of functions, a *concept class*, and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a *loss function*. A learning algorithm (to which $\mathcal{X}, \mathcal{Y}, \mathcal{F}$ and ℓ are known) has access to training data of the form $S = \{(x_i, y_i)\}_{i=1}^{m}$, where (x_i, y_i) are drawn i.i.d. from a probability measure $\mu \in \mathrm{Prob}(\mathcal{X} \times \mathcal{Y})$. Moreover, the learner is given as input a confidence parameter $\delta \in (0,1)$ and an accuracy parameter $\varepsilon \in (0,1)$. Then a learner must output a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ s.t., with probability $\geq 1-\delta$ w.r.t. the choice of training data,

$$\mathbb{E}_{(x,y)\sim\mu}[\ell(y, h(x))] \leq \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}[\ell(y, f(x))] + \varepsilon. \tag{2.1}$$

Note that the first term on the right-hand side vanishes if there exists an $f^* \in \mathcal{F}$ s.t. $\mu(x,y) = \mu_1(x)\delta_{y,f^*(x)}$ $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}$. In this case, we call the learning problem *realizable*, otherwise we refer to it as *agnostic*.

Both in the agnostic and in the realizable scenario, a learning algorithm that always outputs a hypothesis $h \in \mathcal{F}$ is called a *proper learner*, and otherwise it is called *improper*.

A quantity of major interest is the number of examples featuring in such a learning problem. Given a learning algorithm $\mathcal{A}$, the smallest $m = m(\varepsilon,\delta) \in \mathbb{N}$ s.t. the learning requirement (2.1) is satisfied with confidence $1-\delta$ and accuracy ε is called the *sample complexity* of $\mathcal{A}$. The sample complexity of the learning problem is the infimum over the sample complexities of all learning algorithms for the problem. This characterizes, from an information-theoretic perspective, the hardness of a learning problem, but leaves aside questions of computational complexity.

The binary classification problem now arises as a special case from the above if we specify the output space $\mathcal{Y} = \{0,1\}$ and take the loss function to be $\ell(y,\tilde{y}) = 1-\delta_{y,\tilde{y}}$, the 0-1-loss. This setting is well studied and a characterization of its sample complexity is known. At its core is the following combinatorial parameter:

Definition 1 (VC-Dimension Vapnik and Chervonenkis (1971)) Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$. A set $S = \{x_1, \ldots, x_n\} \subset X$ is said to be shattered by $\mathcal{F}$ if for every $b \in \{0,1\}^n$ there exists $f_b \in \mathcal{F}$ s.t. $f_b(x_i) = b_i$ for all $1 \leq i \leq n$.

The Vapnik-Chervonenkis (VC) dimension of $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ is defined to be

$$\mathrm{VCdim}(\mathcal{F}) := \sup\{n \in \mathbb{N}_0 \mid \exists S \subset X \text{ s.t. } |S| = n \text{ and } S \text{ is shattered by } \mathcal{F}\}.$$

The main insight of VC-theory lies in the fact that learnability of a $\{0,1\}$-valued concept class is equivalent to finiteness of its VC-dimension. Even more, the sample

complexity can be expressed in terms of the VC-dimension. This is the content of the following

Theorem 1 (see, e.g., Blumer et al. 1989; Hanneke 2016; Shalev-Shwartz and Ben-David 2014; Vershynin 2018)

In the realizable scenario, the sample complexity of binary classification for a function class $\mathscr{F}$ of VC-dimension d is $m = m(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon}(d + \log 1/\delta)\right)$.

In the agnostic scenario, the sample complexity of binary classification for a function class $\mathscr{F}$ of VC-dimension d is $m = m(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon^2}(d + \log 1/\delta)\right)$.

The proof of the sample complexity upper bound in the agnostic case typically goes via a different complexity measure, the Rademacher complexity, which is then related to the VC-dimension. As this will reappear later on in our analysis, we also recall this definition here.

Definition 2 (Rademacher Complexity (see Bartlett and Mendelson 2002)) Let Z be some space, $\mathscr{F} \subseteq \mathbb{R}^Z$, $z \in Z^n$. The empirical Rademacher complexity of $\mathscr{F}$ w.r.t. z is

$$\hat{\mathscr{R}}(\mathscr{F}) := \mathop{\mathbb{E}}_{\sigma\sim U(\{-1,1\}^n)}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right] = \mathop{\mathbb{E}}_{\sigma\sim U(\{-1,1\}^n)}\left[\sup_{f\in\mathscr{F}}\frac{1}{n}\langle\sigma, f(z)\rangle\right],$$

where $U(\{-1, 1\}^n)$ denotes the uniform distribution on $\{-1, 1\}^n$.

If we consider n i.i.d. random variables $Z_1, ..., Z_n$ distributed according to a probability measure μ on $\mathcal{Z}$ and write $Z = (Z_1, ..., Z_n)$, the Rademacher complexities of $\mathscr{F}$ w.r.t. μ are defined to be $\mathscr{R}_n(\mathscr{F}) := \mathbb{E}_{Z\sim\mu^n}\left[\hat{\mathscr{R}}_{\mathscr{F}}\right]$, $n \in \mathbb{N}$.

3 The binary classification problem with classical instances and quantum labels

We introduce a generalization of the classical binary classification problem to the quantum realm by allowing the two labels to be quantum states. Thus let $\sigma_0, \sigma_1 \in \mathscr{S}(\mathbb{C}^n)$ be two (possibly mixed) quantum states, write $\mathscr{D} = \{\sigma_0, \sigma_1\}$. We assume that classical descriptions of these states (their density matrices) are known to the learning algorithm as well as the fact that only these two quantum labels appear. The class to be learned is now a class of functions $\mathscr{F} \subset \{f : \mathscr{X} \to \mathscr{D}\}$ and the underlying distribution will be a $\mu \in \mathrm{Prob}(\mathscr{X} \times \mathscr{D})$, where $\mathscr{X}$ is some space of classical objects.

We now deviate from the standard PAC setting: We assume the training data to be $S = \{(x_i, \rho_i)\}_{i=1}^m$, $m \in \mathbb{N}$, where the (x_i, ρ_i) are drawn independently according to μ (in particular, $\rho_i \in \mathscr{D}$ for all i). Here, the ρ_i are the actual quantum states, not classical descriptions of them. Therefore, our learning problem is not a classical one, we have to perform measurements on the quantum labels to extract information from them. Equivalently, we represent an example (x_i, ρ_i) drawn from μ as the classical-quantum state

$$\sum_{x,\rho}\mu(x, \rho)|x\rangle\langle x| \otimes \rho,$$

with $\{|x\rangle\}_{x\in\mathscr{X}}$ orthonormal.

Note that this model for the training data differs from the one introduced by Bshouty and Jackson (1998), where the training data consists of copies of a superposition state. Instead, here we assume copies of a mixture of states. This is done mainly for two reasons: First, it allows us to naturally talk about maps with mixed state outputs. Second, it is debatable whether assuming access to superposition examples as in Bshouty and Jackson (1998) is justified (see, e.g., Ciliberto et al. 2018, section 5), and this problem remains when considering maps with quantum outputs. In contrast, the mixtures assumed in our model arise naturally as statistical ensembles of outputs of state preparation procedures, if the parameters of the preparation are chosen according to some (unknown) distribution. In that sense, the form of classical-quantum training data assumed here is both a straightforward generalization of classical training data, given the standard probabilistic interpretation of mixed states, and can (at least in the realizable scenario) be easily imagined to be obtained as outcome of multiple runs of a state preparation experiment with different parameter settings.

A quantum learner for $\mathscr{F}$ with confidence $1 - \delta$ and accuracy ε from $m = m(\varepsilon, \delta)$ quantum examples has to output, for every $\mu \in \mathrm{Prob}(\mathscr{X} \times \mathscr{D})$, with probability $\geq 1 - \delta$ over the choice of training data of size m according to μ, a hypothesis $h \in \mathscr{D}^{\mathscr{X}}$ s.t. $R_\mu(h) \leq \inf_{f\in\mathscr{F}} R_\mu(f) + \varepsilon$. As before, we can consider agnostic versus realizable and proper versus improper variants of this learning model.

Here, we define the risk of a hypothesis $h \in \mathscr{F}$ w.r.t. a distribution $\mu \in \mathrm{Prob}(\mathscr{X} \times \mathscr{D})$ as

$$R_\mu(h) := \int_{\mathscr{X}\times\mathscr{D}} \frac{1}{2}\|\rho - h(x)\|_1 \, \mathrm{d}\mu(x, \rho),$$

where $\|\rho - \sigma\|_1 = \mathrm{tr}[|\rho - \sigma|] = \mathrm{tr}[\sqrt{(\rho - \sigma)^*(\rho - \sigma)}]$ is the Schatten 1-norm.

Note that our assumption on $\mathscr{F}$ implies that $h(x) \in \mathscr{D}\ \forall x \in \mathscr{X}$ and therefore we can easily rewrite

$$R_\mu(h) = \frac{\|\sigma_0 - \sigma_1\|_1}{2} \mathbb{P}_{(x,\rho)\sim\mu}[h(x) \neq \rho],$$

which is just the 0-1-risk multiplied by a constant. We choose the slightly more complicated looking definition for $R_\mu(h)$ for two reasons. On the one hand, $\frac{\|\sigma_0-\sigma_1\|_1}{2}$ is a measure for the distinguishability of σ_0 and σ_1 and thus a natural scale w.r.t. which to measure the prediction error. (Note: If σ_0, σ_1 are orthogonal pure states and thus perfectly distinguishable, the classical scenario is recovered.) On the other hand, our definition of risk can be motivated operationally as we discuss in Appendix 2.

Example 1 Here, we describe a physically motivated problem that is captured by our scenario. The idea is as follows: Suppose we have available a (possibly complicated) ground state preparation procedure. Using this, we want to prepare a ground state $|\varphi_0\rangle$ of a Hamiltonian H. However, H is perturbed by noise about which we have only partial information. We want to learn more about the noise and its influence on the prepared ground state.

We make this idea more concrete. We consider a (self-adjoint) Hamiltonian $H \in \mathbb{C}^{(d+2)\times(d+2)}$ of the form $H = \mathbb{1}_2 \oplus \tilde{H}$, where $\tilde{H} > \mathbb{1}_d$, with (non-unique) ground state $|\varphi_0\rangle := \begin{pmatrix}0 & 1\end{pmatrix}^T \oplus 0$. Suppose that we have a ground state preparation procedure that, if run with Hamiltonian H, prepares $|\varphi_0\rangle$. When implementing this procedure, we have to fix values of a parameter vector $x \in \mathbb{R}^D$. (Think, e.g., of $D = 3$ and x denoting the location at which the experiment is set up.) But due to the laboratory being only imperfectly shielded, there is an unknown region $R \subset \mathbb{R}^D$ in which the system is subject to noise. For simplicity, we assume that only two types of noise can occur and lead to the location-dependent Hamiltonian $H_x^{(i)} = H + \mathbb{1}_{\{x\in R\}} H^{(i)}$, with noise Hamiltonians $H^{(0)} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} \oplus 0$, $H^{(1)} = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \oplus 0$.

The noise can lead to a perturbation of the ground state. Namely:

- For $x \notin R$, $|\varphi_0\rangle$ is a ground state of $H_x^{(i)}$. (This is the case of no effective noise.)
- For $x \in R$, $|\varphi_0\rangle$ is the unique ground state of $H_x^{(0)}$. Hence, the noise $H^{(0)}$ is benign from the perspective of ground state preparation.
- For $x \in R$, $|\varphi_1\rangle := \frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\end{pmatrix}^T \oplus 0$ is the unique ground state of $H_x^{(1)}$. Hence, the noise $H^{(1)}$ is malicious from the perspective of ground state preparation.

Thus, we describe the ground state preparation by a function $f_R^{(i)} : \mathbb{R}^D \to \{|\varphi_0\rangle\langle\varphi_0|, |\varphi_1\rangle\langle\varphi_1|\}$, $f_R^{(i)}(x) = \mathbb{1}_{\{x\notin R\}}|\varphi_0\rangle\langle\varphi_0| + \mathbb{1}_{\{x\in R\}}|\varphi_i\rangle\langle\varphi_i|$. With this formulation, gaining information about the noise region R and the noise type i can be phrased as the problem of (PAC-)learning an unknown element of the (known) function class $\mathscr{F} = \left\{f_R^{(i)}\right\}_{i=0,1,\ R\in\mathscr{R}} \subseteq \{|\varphi_0\rangle\langle\varphi_0|, |\varphi_1\rangle\langle\varphi_1|\}^{\mathbb{R}^D}$, where $\mathscr{R}$ is the class of possible error regions.

Note that $|\varphi_0\rangle$ and $|\varphi_1\rangle$ are not orthogonal and thus cannot be perfectly distinguished. Therefore, we cannot phrase the learning problem as one of binary classification with classical labels.

We return to this setting in Examples 2 and 3 to illustrate our learning strategies.

We want to conclude this section by discussing a drawback of our model. We assume $\mathscr{F} \subset \mathscr{D}^{\mathscr{X}}$, i.e., outputs of any $f \in \mathscr{F}$ are either σ_0 or σ_1. Considering the convex structure of the set of quantum states, which is intimately tied to the probabilistic interpretation of quantum theory, this restriction can be considered unnatural. We nevertheless make it, for two reasons: First, it is easy to show using a Bayesian predictor that, under the assumption of μ being supported on $\mathscr{D}$ (which could, of course, also be contested), the optimal choice of predictors among all functions $(\mathscr{S}(\mathbb{C}^d))^{\mathscr{X}}$ is actually a function in $\mathscr{D}^{\mathscr{X}}$. Second, it is the most direct analog of the classical scenario with binary labels and we consider it a sensible first step that, as demonstrated in Example 1, can already be of physical relevance.

4 Sample complexity upper bounds

4.1 The agnostic case

Our learning strategy is motivated by interpreting the classical training data arising from performing a measurement on the label states as noisy version of the true training data. Before describing the learning strategy, we recall our assumption that classical descriptions of the label states σ_0, σ_1 are known to the learner. Based on this knowledge, the learner can derive the optimal measurement $\{E_0, E_1\}$ for minimum error distinction between the two states, the so-called Holevo-Helstrom measurement (see Watrous 2018, Theorem 3.4), by choosing E_0 to be the orthogonal projector onto the eigenspaces of $\sigma_0 - \sigma_1$ corresponding to nonnegative eigenvalues. This step is where knowledge of the states σ_0 and σ_1 is used.

The learning strategy is now the following, in which we use the Holevo-Helstrom measurement to produce classical training data and thus obtain a classical learning problem:

Noise-corrected Holevo-Helstrom strategy

Given: Quantum training data $S = \{(x_i, \rho_i)\}_{i=1}^m$
Output: Hypothesis $\hat{h} : \mathscr{X} \to \mathscr{D}$
Algorithm:

1. For each i: Perform a Holevo-Helstrom measurement on ρ_i. Let
$$y_i = \begin{cases} 1 & \text{if } E_1 \text{ is accepted} \\ 0 & \text{if } E_1 \text{ is rejected} \end{cases}.$$
2. Let $\tilde{S} = \{(x_i, y_i)\}_{i=1}^m \in (\mathscr{X} \times \{0,1\})^m$. Then one can view (x_i, y_i) as being drawn independently according to the probability measure ν on $\mathscr{X} \times \{0,1\}$ which has
$$\nu_1(x) = \mu_1(x) = \mu(x, \sigma_0) + \mu(x, \sigma_1)$$
as first marginal and
$$\nu(y|x) = \delta_{y0}\left(\mu(\sigma_1|x)\mathrm{tr}[\sigma_1 E_0] + \mu(\sigma_0|x)\mathrm{tr}[\sigma_0 E_0]\right) + \delta_{y1}\left(\mu(\sigma_1|x)\mathrm{tr}[\sigma_1 E_1] + \mu(\sigma_0|x)\mathrm{tr}[\sigma_0 E_1]\right).$$
as the conditional probability distribution of y given x.
3. Use a classical learning algorithm to find $\hat{g} \in \tilde{\mathscr{F}} := \{\tilde{f} : \mathscr{X} \to \{0,1\} \mid \exists f \in \mathscr{F} : f(x) = \sigma_{\tilde{f}(x)} \ \forall x \in \mathscr{X}\}$ s.t. $\tilde{R}_\nu(\hat{g}) := \mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(y, \hat{g}(x)]$ is minimized over $\tilde{\mathscr{F}}$, where
$$\tilde{\ell}(y_1, y_2) := \frac{(1 - \eta_{1\oplus y_2})\mathbb{1}_{y_1 \neq y_2} - \eta_{y_2}\mathbb{1}_{y_1 = y_2}}{1 - \eta_0 - \eta_1},$$
with $\eta_0 = \mathrm{tr}[\sigma_0 E_1]$ and $\eta_1 = \mathrm{tr}[\sigma_1 E_0]$. Here, $\oplus$ denotes addition modulo 2.
4. Define $\hat{h} : \mathscr{X} \to \mathscr{D}$ via $\hat{h}(x) = \sigma_{\hat{g}(x)}$ and output $\hat{h}$ as hypothesis.

Note that the only non-classical step in the strategy is step (1), which consists only of performing local two-outcome measurements.

The modification of the loss function in step (3) gives an unbiased estimate of the true risk:

Lemma 1 (see Natarajan et al. 2013, Lemma 1)
Fix $x \in \mathscr{X}$. With the notation introduced above, for every $z \in \{0,1\}$ it holds that

$$\mathbb{E}_{Y\sim\nu(\cdot|x)}[\tilde{\ell}(z, Y)] = \mathbb{E}_{Y\sim\mu(\cdot|x)}[\mathbb{1}_{z\neq Y}].$$

We can use a standard generalization bound in terms of Rademacher complexities (see, e.g., Theorem 26.5 of Shalev-Shwartz and Ben-David (2014)) to obtain: With probability $\geq 1 - \delta$ over the choice of training data $S = \{(x_i, y_i)\}_{i=1}^m$ according to ν, we have that for all $\tilde{f}^* \in \tilde{\mathscr{F}}$

$$\mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(\hat{g}(x), y)] - \mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(\tilde{f}^*(x), y)] \leq 2\hat{\mathscr{R}}(\tilde{\mathcal{G}}) + \frac{5}{1 - \eta_0 - \eta_1}\sqrt{\frac{2\ln 8/\delta}{m}},$$

where we used that $|\tilde{\ell}(y_1, y_2)| \leq \frac{1}{1-\eta_0-\eta_1}$ and defined the function class

$$\tilde{\mathcal{G}} := \{\mathscr{X} \times \{0,1\} \ni (x, y) \mapsto \tilde{\ell}(\tilde{f}(x), y) \mid \tilde{f} \in \tilde{\mathscr{F}}\}.$$

Next, we relate the empirical Rademacher complexity of $\tilde{\mathcal{G}}$ to that of $\tilde{\mathscr{F}}$.

Lemma 2 *For any training data set $S = \{(x_i, y_i)\}_{i=1}^m$, viewed as an element of $(\mathscr{X} \times \{0,1\})^m$, we have*

$$\hat{\mathscr{R}}(\tilde{\mathcal{G}}) \leq \frac{2}{1 - \eta_0 - \eta_1}\hat{\mathscr{R}}(\tilde{\mathscr{F}}).$$

Proof (Sketch) The proof uses some standard steps that are typically used for example in proving the Lipschitz contraction property of the Rademacher complexity and in studying the Rademacher complexity in a binary classification scenario.

See Appendix 1 for a detailed proof. □

With this, we now reformulate the above result in terms of the VC-dimension. Suppose $\mathrm{VCdim}(\tilde{\mathscr{F}}) = d < \infty$. Then $\hat{\mathscr{R}}(\tilde{\mathscr{F}}) \leq 31\sqrt{\frac{d}{m}}$ (see, e.g., Vershynin 2018, Theorem 8.3.23). Therefore, we obtain that, with probability $\geq 1 - \delta$ over the choice of training data $S = \{(x_i, y_i)\}_{i=1}^m$ according to ν,

$$\mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(\hat{g}(x), y)] - \inf_{\tilde{f}\in\tilde{\mathscr{F}}} \mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(\tilde{f}(x), y)] \leq \frac{124}{1 - \eta_0 - \eta_1}\sqrt{\frac{d}{m}} + \frac{5}{1 - \eta_0 - \eta_1}\sqrt{\frac{2\ln 8/\delta}{m}}.$$

Note that, using Lemma 1, we can now bound

$$\begin{aligned} & R_\mu(\hat{h}) - \inf_{f\in\mathscr{F}} R_\mu(f) \\ &= \frac{\|\sigma_0 - \sigma_1\|_1}{2} \underbrace{\mathbb{E}_{(x,\rho)\sim\mu}[\mathbb{1}_{\hat{g}(x)\neq\rho}]}_{=\mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(\hat{g}(x),y)]} \\ &\quad - \inf_{\tilde{f}\in\tilde{\mathscr{F}}} \frac{\|\sigma_0 - \sigma_1\|_1}{2} \underbrace{\mathbb{E}_{(x,\rho)\sim\mu}[\mathbb{1}_{\tilde{f}(x)\neq\rho}]}_{=\mathbb{E}_{(x,y)\sim\nu}[\tilde{\ell}(\tilde{f}(x),y)]} \\ &\leq \frac{\|\sigma_0 - \sigma_1\|_1}{2}\left(\frac{124}{1 - \eta_0 - \eta_1}\sqrt{\frac{d}{m}} + \frac{5}{1 - \eta_0 - \eta_1}\sqrt{\frac{2\ln 8/\delta}{m}}\right). \end{aligned}$$

Now we can set this equal to ε and rearrange to conclude that a sample size of

$$m \geq \frac{\|\sigma_0 - \sigma_1\|_1^2}{4\varepsilon^2}\left(\frac{124}{1 - \eta_0 - \eta_1}\sqrt{d} + \frac{5}{1 - \eta_0 - \eta_1}\sqrt{2\ln 8/\delta}\right)^2$$

suffices to guarantee that, with probability $\geq 1-\delta$, $R_\mu(\hat{h}) - \inf_{f\in\mathscr{F}} R_\mu(f) \leq \varepsilon$.

If we now observe that $\frac{1}{1-\eta_0-\eta_1} \leq \frac{4}{\|\sigma_0-\sigma_1\|_1}$, we obtain the sample complexity upper bound

$$m = m(\varepsilon,\delta) = \mathscr{O}\left(\frac{d}{\varepsilon^2} + \frac{\log 1/\delta}{\varepsilon^2}\right).$$

Remark 1 The naive version of our learning strategy would be to perform Holevo-Helstrom measurements and then apply a classical learning strategy, like empirical risk minimization, without correcting for the noise in the resulting classical labels. Actually, this learning strategy already performs reasonably well and, in certain special cases, even allows to reduce the quantum learning problem to a fully classical one. For a detailed analysis of the performance of this simpler strategy, the reader is referred to Appendix 3.

Example 2 We illustrate our agnostic learning strategy for the scenario of Example 1. As discussed in Appendix 3, as both label states $|\varphi_0\rangle\langle\varphi_0|$ and $|\varphi_1\rangle\langle\varphi_1|$ are pure, we can actually dispense with the modification of the classical loss function and simply take the 0-1-loss. Therefore, the Holevo-Helstrom strategy will look as follows: We first perform local Holevo-Helstrom measurements with measurement operators $E_0 \propto \left(-1+\sqrt{2}\ 1\right)^T \left(-1+\sqrt{2}\ 1\right) \oplus 0$, $E_1 = \mathbb{1}_{2+d} - E_0$. This gives rise to classical training data. With that data, we then perform (classical) empirical risk minimization over the class $\tilde{\mathscr{F}} = \left\{\tilde{f}_R^{(i)}\right\}_{i=0,1,\ R\in\mathscr{R}}$, where $\tilde{f}_R^{(i)} : \mathbb{R}^D \to \{0,1\}$, $\tilde{f}_R^{(i)} : \mathbb{R}^D \to \{0,1\}$, $\tilde{f}_R^{(i)}(x) = \mathbb{1}_{\{x\in R\}}\delta_{i,1}$. Note that $f_R^{(0)}$ is the zero-function for every $R \in \mathscr{R}$.

Both the optimization procedure and the generalization capability depend on the class $\mathscr{R}$ of possible noise regions. Concerning the generalization performance, observerve that, if $\emptyset \in \mathscr{R}$, then $\mathrm{VCdim}(\tilde{\mathscr{F}}) = \mathrm{VCdim}(\tilde{\mathscr{F}}_{\mathscr{R}})$, where we take $\tilde{\mathscr{F}}_{\mathscr{R}} = \{\mathbb{R}^D \ni x \mapsto \mathbb{1}_{\{x\in R\}} \mid R \in \mathscr{R}\}$ to be the class of indicator functions of sets from $\mathscr{R}$. The VC-dimension of such classes is well known for different geometric classes $\mathscr{R}$. E.g., if $\mathscr{R}$ is the class of axis-aligned rectangles or that of Euclidean balls in $\mathbb{R}^D$, then $\mathrm{VCdim}(\tilde{\mathscr{F}}_{\mathscr{R}})$ scales linearly in D and thus the dependence of the sample complexity upper bound on the number of parameters D is linear. If, however, we take $\mathscr{R}$ to be the class of compact and convex subsets of $\mathbb{R}^D$, then $\mathrm{VCdim}(\tilde{\mathscr{F}}_{\mathscr{R}}) = \infty$ and the sample complexity upper bound becomes void. This is congruent with the intuition that without prior assumptions on the structure of the regions that can be influenced by noise, learning the noise (in particular its region) will be hard and maybe infeasible.

4.2 The realizable case

The strategy from the previous subsection uses a generalization bound via the Rademacher complexity and yields a sample complexity bound depending quadratically on $1/\varepsilon$. In the classical binary classification problem it is known (see Theorem 1) that under the realizability assumption this can be improved to $1/\varepsilon$, but this typically requires a different kind of reasoning via ε-nets. (Compare section 28.3 of Shalev-Shwartz and Ben-David (2014)). In Theorem 6 we show how the reasoning by Hanneke (2016) can be combined with results by Laird (1988) to achieve the $1/\varepsilon$-scaling also in the case of two-sided classification noise. This sample complexity upper bound is seen to be optimal in its dependence on the VC-dimension d, the error rate bound η, the confidence δ and the accuracy ε by a comparison to the lower bound in Theorem 27 of Arunachalam and de Wolf (2018).

If, as in the previous subsection, we consider the classical training data obtained by measuring the quantum training data as noisy version of a true sample, we can exchange step 3 in the Holevo-Helstrom strategy by the minimum disagreement-based classical learning strategy achieving the optimal sample complexity bound of Theorem D.2. This directly yields the following

Theorem 2 *Let $\sigma_0, \sigma_1 \in \mathscr{S}(\mathbb{C}^n)$ be (distinct) quantum states. Let $\varepsilon \in (0,1)$, $\delta \in (0, 2\cdot(\frac{2e}{d})^d)$, where d is the VC-dimension of $\mathscr{F} \subset \{0,1\}^{\mathscr{X}}$. Then*

$$m = \mathscr{O}\left(\frac{1}{\varepsilon(1-2\max\{\mathrm{tr}[E_0\sigma_1], \mathrm{tr}[E_1\sigma_0]\})^2}\,(d+\log 1/\delta)\right)$$

quantum examples of a function in $\mathscr{F}$ are sufficient for binary classification with classical instances and quantum labels σ_0, σ_1 with accuracy ε and confidence $1-\delta$.

Example 3 When considering this learning strategy in the setting of Example 1, we first perform the Holevo-Helstrom measurements as in Example 2 to obtain classical data. Again, this is followed by a classical learning procedure for the class $\tilde{\mathscr{F}} = \left\{\tilde{f}_R^{(i)}\right\}_{i=0,1,\ R\in\mathscr{R}}$.

Whereas the sample complexity bound derived for the agnostic case in Section 4.1 applies to any (noise-corrected) classical empirical risk minimization, the procedure leading to the bound in Theorem 2 is a specific one, presented in the proof of Theorem D.2. First, the classical data is processed, using the subsampling algorithm of Hanneke (2016) (see Algorithm 2), to generate a collection of subsamples. For each of those subsamples, we then apply Algorithm 1: We use a first part of the subsample to group the elements of $\tilde{\mathscr{F}}$ into equivalence classes (according how they act on that part of the subsample), and the remainder is used to test

the performance of each equivalence class. Afterwards, we output as hypothesis for that subsample a representative of the equivalence class that performs best in that test, i.e., that minimizes the number of disagreements with the part of the subsample used for testing. Whether and how the grouping into equivalence classes and finding minimum disagreement strategies can be done (efficiently) depends on $\tilde{\mathscr{F}}$, and thus on $\mathscr{R}$. Finally, we take a majority vote over all the subsample hypotheses to get the output hypothesis of the classical learning procedure.

The dependence of the sample complexity on $\tilde{\mathscr{F}}$ via the VC-dimension of the class of indicator functions of sets from $\mathscr{R}$ is analogous to Example 2.

Remark 2 From the description of our noise-corrected Holevo-Helstrom strategy (either in the form of Section 4.1 or that of this subsection), we can directly see that whether it is a proper or an improper learner depends on whether the classical learning algorithm in step (3) is. As the classical learning algorithm used in Section 4.1 is a simple Empirical Risk Minimization, it is in particular proper. So our noise-corrected Holevo-Helstrom strategy for the agnostic case is proper as well. The classical learner used in this subsection, however, is in general improper. So also the noise-corrected Holevo-Helstrom strategy for the realizable case is in general improper.

5 Sample complexity lower bounds

Whereas the goal of the previous section was to give strategies for solving the binary classification problem with classical instances and quantum labels and to prove upper bounds on the sufficient number of classical-quantum examples, we now turn to the complementary question of lower bounds on the number of required examples. In this section, we derive lower bounds that match the respective upper bounds from the previous section, and therefore, we conclude that the procedures described in Section 4 are optimal w.r.t. sample size in terms of the dependence on ε, δ, and d.

5.1 The agnostic case

We prove the sample complexity lower bounds in two parts, the first depending on the confidence parameter δ but not on the VC-dimension of the function class and conversely for the second.

We establish the VC-dimension-independent sample complexity lower bound in the following

Lemma 3 *Let $\sigma_0, \sigma_1 \in \mathscr{S}(\mathbb{C}^n)$, let $\varepsilon \in (0, \frac{\|\sigma_0-\sigma_1\|_1}{2\sqrt{2}})$, $\delta \in (0,1)$. Let $\mathscr{F} \subset \mathscr{D}^{\mathscr{X}}$ be a non-trivial concept class. Suppose $\mathscr{A}$ is a learning algorithm that solves the binary classification task with classical instances and (distinct) label states σ_0, σ_1 and concept class $\mathscr{F}$ with confidence $1-\delta$ and accuracy ε using $m = m(\varepsilon, \delta)$ examples. Then $m \geq \Omega\left(\|\sigma_0 - \sigma_1\|_1^2 \frac{\log 1/\delta}{\varepsilon^2}\right)$.*

Proof (Sketch) As $\mathscr{F}$ is non-trivial, there exist concepts $f, g \in \mathscr{F}$ and a point $x \in \mathscr{X}$ s.t. $f(x) = \sigma_0$ and $g(x) = \sigma_1$. Let $\lambda = \frac{\varepsilon}{2\|\sigma_0-\sigma_1\|_1} \in (0,1)$. Define probability distributions $\mu_\pm$ on $\mathscr{X} \times \mathscr{D}$ via

$$\mu_\pm(x, f(x)) = \frac{1 \pm \lambda}{2}, \quad \mu_\pm(x, g(x)) = \frac{1 \mp \lambda}{2}.$$

By explicitly evaluating the risk $R_\pm(h)$, we see that achieving an excess risk $\leq \varepsilon$ with probability $\geq 1-\delta$, requires the learner to distinguish between the underlying distributions $\mu_\pm$, and thus the corresponding training data states $\rho_\pm^{\otimes m}$, with probability $\geq 1-\delta$.

It is well known (see, e.g., Nielsen and Chuang 2009, chapter 9) that the optimal success probability of this quantum distinguishing task is given by

$$p_{\text{opt}} = \frac{1}{2}(1 + \frac{1}{2}\left\|\rho_+^{\otimes m} - \rho_-^{\otimes m}\right\|_1).$$

Via the Fuchs-van de Graaf inequalities, which state that

$$\frac{1}{2}\left\|\rho_1^{\otimes m} - \rho_2^{\otimes m}\right\|_1 \leq \sqrt{1 - F(\rho_1^{\otimes m}, \rho_2^{\otimes m})^2} = \sqrt{1 - F(\rho_1, \rho_2)^{2m}},$$

this can be upper bounded using lower bounds on the fidelity $F(\rho_+^{\otimes m}, \rho_-^{\otimes m}) = F(\rho_+, \rho_-)^m$. The fidelity $F(\rho_+, \rho_-)$ can be lower-bounded using its strong concavity and the explicit expressions for $\rho_\pm$. The result then follows by comparing the obtained upper bound with the required lower bound $p_{\text{opt}} \geq 1-\delta$.

See Appendix 1 for a detailed proof. □

For the proof of the VC-dimension-dependent part of the lower bound we need a well known observation about the eigenvalues of a statistical mixture of two pure quantum states, which is the content of the following

Lemma 4 *Let $|\psi\rangle, |\phi\rangle \in \mathbb{C}^n$ be distinct pure quantum states. Let $\alpha, \beta \geq 0$ be real numbers. Then the non-zero eigenvalues of the mixture $\rho := \alpha|\psi\rangle\langle\psi| + \beta|\phi\rangle\langle\phi|$ are given by*

$$\lambda_{1/2}(\rho) = \frac{\alpha + \beta \pm \sqrt{(\alpha-\beta)^2 + 4\alpha\beta|\langle\psi|\phi\rangle|^2}}{2}.$$

With this we can now prove a sample complexity lower bound for the case of pure label states.

Theorem 3 *Let $\sigma_0 = |\psi_0\rangle\langle\psi_0|, \sigma_1 = |\psi_1\rangle\langle\psi_1| \in \mathscr{S}(\mathbb{C}^n)$ be (distinct) pure quantum states, let $\varepsilon \in (0, \frac{\|\sigma_0-\sigma_1\|_1}{8})$,*

$\delta \in (0, 1 - H\left(\frac{1}{4}\right))$. *Let* $\mathscr{F} \subset \mathscr{D}^{\mathscr{X}}$ *be a non-trivial concept class s.t.* $\tilde{\mathscr{F}}$ *has VC-dimension* d. *Suppose* $\mathscr{A}$ *is a learning algorithm that solves the binary classification task with classical instances and (distinct) label states* σ_0, σ_1 *and concept class* $\mathscr{F}$ *with confidence* $1-\delta$ *and accuracy* ε *using* $m = m(\varepsilon, \delta)$ *examples. Then* $m \geq \Omega\left(\frac{d}{\varepsilon^2}\right)$.

Proof (Sketch) We follow the information-theoretic proof strategy from Arunachalam and de Wolf (2018). Let $S = (s_1, \ldots, s_d) \in \mathscr{X}$ be a set shattered by $\tilde{\mathscr{F}}$, for each $a \in \{0,1\}^d$ define the distribution μ_a on $\{1, \ldots, d\} \times \{0,1\}$ via

$$\mu_a(i,b) := \frac{1}{2d}\left(1 + (-1)^{a_i + b}\frac{8\varepsilon}{\|\sigma_0 - \sigma_1\|_1}\right).$$

Note that $\forall a \in \{0,1\}^d \; \exists f_a \in \tilde{\mathscr{F}} : f_a(s_i) = a_i$ by shattering and that f_a is a minimum error concept w.r.t. μ_a. By evaluating the excess error of an $f_{\tilde{a}}$ compared to f_a, we see that solving the learning problem with confidence $1-\delta$ requires the learner to output, with probability $\geq 1-\delta$, a hypothesis described by a string whose Hamming distance to the true underlying string is $\leq \frac{d}{4}$. We can use this observation to obtain the lower bound $I(A:B) \geq \Omega(d)$ on the mutual information between underlying string A (drawn uniformly at random) and corresponding quantum training data B.

We can also upper bound the mutual information. A standard argument shows $I(A:B) \leq m \cdot I(A:B_1)$, where m is the number of copies of the quantum example state and B_1 describes a single quantum example state. Using Lemma 4 and the explicit expression for a quantum example state, we can compute $I(A:B_1)$ and use Taylor expansion to see that $I(A:B_1) \leq \mathcal{O}(\varepsilon^2)$. Comparing the lower and upper bounds on $I(A:B)$ now gives $m \geq \Omega\left(\frac{d}{\varepsilon^2}\right)$.

See Appendix 1 for a detailed proof. □

If we now combine Lemma 3 and Theorem 3 with the result of Section 4.1 we obtain

Corollary 1 *Let* $\sigma_0, \sigma_1 \in \mathscr{S}(\mathbb{C}^n)$ *be (distinct) pure quantum states, let* $\varepsilon \in (0, \frac{\|\sigma_0 - \sigma_1\|_1}{8})$, $\delta \in (0, 1 - H\left(\frac{1}{4}\right))$. *Let* $\mathscr{F} \subset \mathscr{D}^{\mathscr{X}}$ *be a non-trivial concept class s.t.* $\tilde{\mathscr{F}}$ *has VC-dimension* d. *Then a sample size of* $\Theta\left(\frac{d}{\varepsilon^2} + \frac{\log 1/\delta}{\varepsilon^2}\right)$ *is necessary and sufficient for solving the binary classification task with classical instances and quantum labels* σ_0, σ_1 *and hypothesis class* $\mathscr{F}$ *with confidence* $1-\delta$ *and accuracy* ε.

Therefore, we have shown that the strategy from Section 4.1 is, for pure states, optimal in sample complexity w.r.t. its dependence the VC-dimension, the accuracy and the confidence. But we do not make a statement on optimality w.r.t. the dependence on the distinguishability of the label states, because the parameter $\|\sigma_0 - \sigma_1\|_1$ is lacking from our lower bound.

5.2 The realizable case

We now show analogous lower bounds for the sample complexity in the realizable scenario with the same proof strategy.

Lemma 5 *Let* $\sigma_0, \sigma_1 \in \mathscr{S}(\mathbb{C}^n)$, *let* $\varepsilon \in (0, \frac{\|\sigma_0 - \sigma_1\|_1}{2})$, $\delta \in (0, \frac{1}{2})$. *Let* $\mathscr{F} \subset \mathscr{D}^{\mathscr{X}}$ *be a non-trivial concept class. Suppose* $\mathscr{A}$ *is a learning algorithm which solves the binary classification task with classical instances and (distinct) label states* σ_0, σ_1 *and concept class* $\mathscr{F}$ *with confidence* $1-\delta$ *and accuracy* ε *using* $m = m(\varepsilon, \delta)$ *examples in the realizable scenario. Then* $m \geq \Omega\left(\frac{\log 1/\delta}{\varepsilon}\right)$.

Proof This can be proved similarly to Lemma 3. See Appendix 1 for a detailed proof. □

We now provide the analog of Theorem 3 for the realizable case.

Theorem 4 *Let* $\sigma_0 = |\psi_0\rangle\langle\psi_0|, \sigma_1 = |\psi_1\rangle\langle\psi_1| \in \mathscr{S}(\mathbb{C}^n)$ *be (distinct) pure quantum states, let* $\varepsilon \in (0, \frac{\|\sigma_0 - \sigma_1\|_1}{8})$, $\delta \in (0, \frac{1}{2})$. *Let* $\mathscr{F} \subset \mathscr{D}^{\mathscr{X}}$ *be a non-trivial concept class s.t.* $\tilde{\mathscr{F}}$ *has VC-dimension* $d+1$. *Suppose* $\mathscr{A}$ *is a learning algorithm which solves the binary classification task with classical instances and (distinct) label states* σ_0, σ_1 *and concept class* $\mathscr{F}$ *with confidence* $1-\delta$ *and accuracy* ε *using* $m = m(\varepsilon, \delta)$ *examples in the realizable case. Then* $m \geq \Omega\left(\frac{d}{\varepsilon}\right)$.

Proof This can be proved similarly to Theorem 3. See Appendix 1 for a detailed proof. □

Thus, we have obtained a sample complexity lower bound that matches the upper bound proved in Section 4.2 in the dependence on the VC-dimension, the confidence and the accuracy, but we do not make a statement about optimality w.r.t. the dependence on $\|\sigma_0 - \sigma_1\|_1$.

Remark 3 As already discussed in Section 2.1, in proving the sample complexity lower bounds we resort to the Heisenberg picture, which allows us to absorb the intermediate quantum channels performed by a learner into the measurement. These lower bounds therefore even hold for quantum learning algorithms that perform coherent and adaptive measurements on the training data. In particular, the information-theoretic complexity of our learning problem does not change if we restrict the quantum learner to only performing two-outcome POVMs locally (i.e., on one subsystem only). This is maybe not too much of a

surprise, since the optimal measurement for distinguishing states drawn uniformly at random from $\{\bigotimes_{i=1}^m \sigma_{x_i}\}_{x\in\{0,1\}^m}$ can, using the Holevo-Yuen-Kennedy-Lax optimality criterion (Holevo 1973; Yuen et al. 1975), be seen to be exactly given by local Holevo-Helstrom measurements.

6 Conclusion and outlook

We have proposed a novel way of modifying the classical binary classification problem to obtain a quantum counterpart. The conceptual difference to the framework of quantum PAC learning as discussed in Arunachalam and de Wolf (2017) is that we work with maps whose outputs are themselves quantum states, not classical labels. This naturally gives rise to training data given by quantum states, which is one aspect in which our setting differs from Aaronson (2007).

Using results from classical learning theory on dealing with classification noise in the training data, we exhibited learning strategies (based on the Holevo-Helstrom measurement) for binary classification with classical instances and quantum labels. The learning strategies consist of two main steps: First, classical information is extracted from the training data by performing a (localized) measurement. Second, classical learning strategies are applied. We complemented these procedures by sample complexity lower bounds thereby establishing the information-theoretic optimality of these strategies for pure label states w.r.t. the dependence on VC-dimension, confidence and accuracy.

We conclude with some open questions that we leave open for further research:

- Can we derive sample complexity lower bounds which explicitly incorporate factors related to the hardness of distinguishing σ_0 and σ_1, e.g., in terms of $\|\sigma_0-\sigma_1\|_1$ or $\max\{\mathrm{tr}[E_0\sigma_1], \mathrm{tr}[E_1\sigma_0]\}$? Or can the corresponding factors in the upper bounds be eliminated? Could this be related to another complexity measure from classical learning theory, the "fat-shattering dimension" of the class

$$\{\mathscr{X}\times\mathscr{E}(\mathbb{C}^d)\ni(x,E)\mapsto \mathrm{tr}[Ef(x)]\mid f\in\mathscr{F}\}?$$

- Our analysis is focused on the information-theoretic part of the learning problem, i.e., the sample complexity. Can we improve the computational complexity?
- For deriving our sample complexity upper bounds, we used specific classical learning procedures applied to the post-measurement training data. In the agnostic case, we use empirical risk minimization, in the realizable case we use a combination of a minimum disagreement approach with a subsampling procedure. In both cases, we decided for these algorithms to achieve the (essentially) optimal sample complexity characterized via the VC-dimension.

  However, we could use other classical learning procedures for "post-processing". Can we identify situations in which procedures like structural risk minimization, compression schemes, or stable learning procedures yield useful sample complexity bounds?
- We considered the case of classical instances. Can this be extended to a scenario of quantum instances with classical (or even quantum) labels? Whereas we were able to study the case of classical instances and quantum labels with methods from learning with label noise, once the instances themselves are quantum, we might have to employ ideas from learning models with restricted access to the instances such as that of "learning with restricted focus of attention" proposed in Ben-David and Dichterman (1998).
- Our strategy uses the Holevo-Helstrom measurement which can be understood as inducing the minimum amount of noise. However, in classical learning theory it is well known that adding noise to the training data can be helpful in preventing overfitting. In this spirit, can we justify other measurements than the Holevo-Helstrom measurement?
- We assumed throughout our analysis that the learning algorithm has to output a hypothesis that maps into $\{\sigma_0,\sigma_1\}$. What if we allow for hypotheses that map into $\mathrm{conv}(\{\sigma_0,\sigma_1\})$ or $\mathscr{S}(\mathbb{C}^d)$?
- Finally, we assume throughout that the label states σ_0, σ_1 are known in advance. Can this assumption be removed? Here, it might be helpful that Theorem 6 does not need explicit knowledge of the error rates η_0, η_1, but merely of an upper bound η_b on them.

Appendix 1. Proofs

Proof of Lemma 2 Let $z = ((x_i,y_i))_{i=1}^m \in (\mathscr{X}\times\{0,1\})^m$. If we use $\mathbb{1}_{\tilde f(x_i)\neq y_i} = \frac{1-(1-2\tilde f(x_i))(1-2y_i)}{2}$ and $\mathbb{1}_{\tilde f(x_i)=y_i} = \frac{1+(1-2\tilde f(x_i))(1-2y_i)}{2}$, then we can rewrite

$$\begin{aligned}\hat{\mathscr{R}}(\tilde{\mathcal{G}}) &= \mathbb{E}_\sigma\Big[\sup_{\tilde f\in\tilde{\mathscr{F}}}\frac{1}{m}\sum_{i=1}^m \sigma_i\tilde\ell(\tilde f(x_i),y_i)\Big]\\ &= \mathbb{E}_\sigma\Bigg[\sup_{\tilde f\in\tilde{\mathscr{F}}}\frac{1}{m}\sum_{i=1}^m\sigma_i\frac{1}{1-\eta_0-\eta_1}\\ &\quad\times\Bigg((1-\eta_{1\oplus y_i})\frac{1-(1-2\tilde f(x_i))(1-2y_i)}{2}\\ &\quad-\eta_{y_i}\frac{1+(1-2\tilde f(x_i))(1-2y_i)}{2}\Bigg)\Bigg].\end{aligned}$$

Next, we use that $\mathbb{E}_\sigma[\sigma_i] = 0$ and that σ_i and $(1-2y_i)\sigma_i$ have the same distribution for all i. With this we obtain from the above

$$\begin{aligned}
\hat{\mathscr{R}}(\tilde{\mathcal{G}}) &= \frac{1}{1-\eta_0-\eta_1}\mathbb{E}_\sigma\left[\sup_{\tilde{f}\in\tilde{\mathscr{F}}}\frac{1}{m}\sum_{i=1}^m \sigma_i(1-\eta_{1\oplus y_i}+\eta_{y_i})\tilde{f}(x_i)\right] \\
&= \frac{1}{2(1-\eta_0-\eta_1)}\mathbb{E}_{\sigma_2,\ldots,\sigma_m}\left[\sup_{\tilde{f},\tilde{f}'\in\tilde{\mathscr{F}}}\frac{1}{m}\underbrace{(1-\eta_{1\oplus y_1}+\eta_{y_1})(\tilde{f}(x_1)-\tilde{f}'(x_1))}_{\leq 2|\tilde{f}(x_1)-\tilde{f}'(x_1)|}\right. \\
&\quad \left. + \frac{1}{m}\sum_{i=2}^m \sigma_i(1-\eta_{1\oplus y_i}+\eta_{y_i})(\tilde{f}(x_i)+\tilde{f}'(x_i))\right] \\
&\leq \frac{1}{1-\eta_0-\eta_1}\mathbb{E}_\sigma\left[\sup_{\tilde{f}\in\tilde{\mathscr{F}}}\frac{2}{m}\sigma_1\tilde{f}(x_1)+\frac{1}{m}\sum_{i=2}^m \sigma_i(1-\eta_{1\oplus y_i}+\eta_{y_i})\tilde{f}(x_i)\right],
\end{aligned}$$

where the last step used that the expression is invariant w.r.t. interchanging $\tilde{f}$ and $\tilde{f}'$, so we can drop the absolute value. Now we can iterate this reasoning for $i = 2, \ldots, m$ and obtain

$$\begin{aligned}
&\hat{\mathscr{R}}(\tilde{\mathcal{G}} \leq \frac{2}{1-\eta_0-\eta_1}\mathbb{E}_\sigma\left[\sup_{\tilde{f}\in\tilde{\mathscr{F}}}\frac{1}{m}\sum_{i=1}^m \sigma_i\tilde{f}(x_i)\right] \\
&= \frac{2}{1-\eta_0-\eta_1}\hat{\mathscr{R}}(\tilde{\mathscr{F}}),
\end{aligned}$$

the desired inequality. □

Proof of Lemma 3 As $\mathscr{F}$ is non-trivial, there exist concepts $f, g \in \mathscr{F}$ and a point $x \in \mathscr{X}$ s.t. $f(x) = \sigma_0$ and $g(x) = \sigma_1$. Let $\lambda \in (0, 1)$ (to be chosen appropriately later in the proof). Define probability distributions $\mu_\pm$ on $\mathscr{X} \times \mathscr{Y}$ via

$$\mu_\pm(x, f(x)) = \frac{1 \pm \lambda}{2}, \quad \mu_\pm(x, g(x)) = \frac{1 \mp \lambda}{2}.$$

The risk of a hypothesis $h \in \mathscr{Y}^{\mathscr{X}}$ w.r.t. these probability measures is given by

$$\begin{aligned}
R_\pm(h) &= \frac{1\pm\lambda}{4}\|\sigma_0 - h(x)\|_1 + \frac{1\mp\lambda}{4}\|\sigma_1 - h(x)\|_1 \\
&= \begin{cases} \frac{1\pm\lambda}{4}\|\sigma_0-\sigma_1\|_1 & \text{if } h(x)=\sigma_1 \\ \frac{1\mp\lambda}{4}\|\sigma_0-\sigma_1\|_1 & \text{if } h(x)=\sigma_0 \end{cases},
\end{aligned}$$

in particular the optimal achievable risk is $\frac{1-\lambda}{4}\|\sigma_0 - \sigma_1\|_1$. Note that a hypothesis which predicts the suboptimal label state for x has an excess risk of

$$\frac{1+\lambda}{4}\|\sigma_0-\sigma_1\| - \frac{1-\lambda}{4}\|\sigma_0-\sigma_1\|_1 = \frac{\lambda}{2}\|\sigma_0-\sigma_1\|_1.$$

So if we pick $\lambda = \frac{\varepsilon}{2\|\sigma_0-\sigma_1\|_1} < 1$, then in order to achieve an excess risk $\leq \varepsilon$ with probability $\geq 1-\delta$, the learning algorithm has to be able to distinguish between the underlying distributions $\mu_\pm$ with probability $\geq 1-\delta$.

As the algorithm has access to the underlying distribution only via the training data, this means that the algorithm has to be able to distinguish the corresponding training data ensembles with probability $\geq 1-\delta$. Here, we observe that the training data being drawn i.i.d. according to $\mu_\pm$ is equivalent to the learning algorithm having access to m copies of the state

$$\rho_\pm := \mu_\pm(x, f(x))|x\rangle\langle x| \otimes \sigma_0 + \mu_\pm(x, g(x))|x\rangle\langle x| \otimes \sigma_1,$$

because this mixed state simply describes the statistical mixture. The optimal success probability for distinguishing between two quantum states is a well-studied object in quantum information theory. It can be characterized by the trace distance between the two states and is given (in our case) by (see, e.g., Nielsen and Chuang 2009)

$$p_{\text{opt}} = \frac{1}{2}(1 + \frac{1}{2}\left\|\rho_+^{\otimes m} - \rho_-^{\otimes m}\right\|_1).$$

As the trace distance of tensor products is not that easy to deal with, we will instead work with the fidelity defined as

$$F(\rho, \sigma) := \text{tr}[\sqrt{\rho^{\frac{1}{2}}\sigma\rho^{\frac{1}{2}}}].$$

According to the Fuchs-van de Graaf inequalities we have

$$\begin{aligned}
&\frac{1}{2}\left\|\rho_+^{\otimes m} - \rho_-^{\otimes m}\right\|_1 \leq \sqrt{1 - F(\rho_+^{\otimes m}, \rho_-^{\otimes m})^2} \\
&= \sqrt{1 - F(\rho_+, \rho_-)^{2m}},
\end{aligned}$$

where the last steps uses multiplicativity of the fidelity under tensor products. Now we require $p_{\text{opt}} \geq 1-\delta$ and rearrange to obtain

$$F(\rho_+, \rho_-)^{2m} \leq 4\delta(1-\delta)$$

or equivalently after taking logarithms

$$m \geq \frac{\log(4\delta(1-\delta))}{\log(F(\rho_+, \rho_-)^2)}.$$

By strong concavity of the fidelity, we have

$$\begin{aligned} F(\rho_+, \rho_-) &\geq \sqrt{\frac{1+\lambda}{2}\frac{1-\lambda}{2}} F(|x\rangle\langle x| \otimes f(x), |x\rangle\langle x| \otimes f(x)) \\ &\quad + \sqrt{\frac{1-\lambda}{2}\frac{1+\lambda}{2}} F(|x\rangle\langle x| \otimes g(x), |x\rangle\langle x| \otimes g(x)) \\ &= \sqrt{1-\lambda^2}. \end{aligned}$$

This now implies

$$m \geq \frac{\log(4\delta(1-\delta))}{\log(F(\rho_+, \rho_-)^2)} = \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\log\left(\frac{1}{F(\rho_+,\rho_-)^2}\right)} \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\log\left(\frac{1}{1-\lambda^2}\right)}.$$

Thus, we obtain (after Taylor-expanding the logarithm in the denominator)

$$m \geq \Omega\left(\|\sigma_0 - \sigma_1\|_1^2 \frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}\right),$$

as desired. □

Proof of Lemma 4 Pick an orthonormal basis $\{|k\rangle\}_{k=1,\ldots,n}$ of $\mathbb{C}^n$ s.t. $|\psi\rangle = |0\rangle$ and $|\phi\rangle = \cos(\varphi)|0\rangle + \sin(\varphi)|1\rangle$ for an angle $0 \leq \varphi < 2\pi$. Then, when restricting to the relevant subspace spanned by $|0\rangle$ and $|1\rangle$, we get

$$\rho|_{\text{span}\{|0\rangle,|1\rangle\}} = \begin{pmatrix} \alpha + \beta\cos^2(\varphi) & \beta\cos(\varphi)\sin(\varphi) \\ \beta\cos(\varphi)\sin(\varphi) & \beta\sin^2(\varphi) \end{pmatrix} =: A.$$

We now easily see that

$$\det(A) = \alpha\beta\sin^2(\varphi) \stackrel{!}{=} \lambda_1\lambda_2 \text{ and } \text{tr}[A] = \alpha + \beta \stackrel{!}{=} \lambda_1 + \lambda_2,$$

where λ_1, λ_2 are the two non-zero eigenvalues of ρ. We can solve the second of these two equations for λ_2 and plug this back into the first equation to obtain

$$\lambda_1^2 - \lambda_1(\alpha+\beta) + \alpha\beta\sin^2(\varphi) = 0.$$

We now solve this quadratic equation and obtain the two eigenvalues

$$\begin{aligned} \lambda_{1/2} &= \frac{\alpha+\beta \pm \sqrt{\alpha^2+\beta^2+2\alpha\beta(2\cos^2(\varphi)-1)}}{2} \\ &= \frac{\alpha+\beta \pm \sqrt{(\alpha-\beta)^2+4\alpha\beta|\langle\psi|\phi\rangle|^2}}{2}, \end{aligned}$$

where we used that $|\cos(\varphi)| = |\langle\psi|\phi\rangle|$. □

Detailed Proof of Theorem 3 Let $S = (s_1, \ldots, s_d) \in \mathcal{X}$ be a set shattered by $\tilde{\mathcal{F}}$, for each $a \in \{0,1\}^d$ define the distribution μ_a on $\{1, \ldots, d\} \times \{0, 1\}$ via

$$\mu_a(i,b) := \frac{1}{2d}\left(1 + (-1)^{a_i+b}\frac{8\varepsilon}{\|\sigma_0 - \sigma_1\|_1}\right).$$

Note that $\forall a \in \{0,1\}^d\ \exists f_a \in \tilde{\mathcal{F}} : f_a(s_i) = a_i$ by shattering and that for each $a \in \{0,1\}^d$, f_a is a minimum error concept w.r.t. μ_a and a concept $f_{\tilde{a}}$ has additional error

$$d_H(a,\tilde{a})\frac{8\varepsilon}{d\|\sigma_0-\sigma_1\|_1} \cdot \frac{\|\sigma_0-\sigma_1\|_1}{2} = d_H(a,\tilde{a})\frac{4\varepsilon}{d}$$

compared to f_a. Hence, in order to solve the learning problem with confidence $1-\delta$ and accuracy ε the algorithm $\mathscr{A}$ has to output, with probability $\geq 1-\delta$, a hypothesis (generated from the training data arising from the underlying string) that when evaluated on S yields a vector that is $\frac{d}{4}$-close to the underlying string in Hamming distance.

Let A be a random variable distributed uniformly on $\{0,1\}^d$ (corresponding to the unknown underlying string a). Let $B = B_1 \ldots B_m$ be the training data with each example generated independently from μ_a described by the quantum ensemble

$$\mathscr{E}_a = \{\mu_a(i,b), |s_i\rangle\langle s_i| \otimes \sigma_b\}_{i=1,\ldots,d,\ b=0,1},$$

or, equivalently, by the quantum state

$$\rho_a = \sum_{i=1}^{d} |s_i\rangle\langle s_i| \otimes (\mu_a(i,0)\sigma_0 + \mu_a(i,1)\sigma_1).$$

In particular, the composite system of underlying string and corresponding training data is described by the quantum state

$$\sigma_{AB} = \frac{1}{2^d}\sum_{a\in\{0,1\}^d} |a\rangle\langle a| \otimes \rho_a^{\otimes m}.$$

We follow the information-theoretic proof strategy from Arunachalam and de Wolf (2018), i.e., we first show a lower bound on the mutual information $I(A:B)$ which arises from the learning requirement, then observe that $I(A:B) \leq m \cdot I(A:B_1)$ and finally upper bound the mutual information $I(A:B_1)$.

First for the mutual information lower bound. Let $h(B) \in \{0,1\}^d$ denote the label vector assigned to S by the hypothesis produced by the learner upon input of training data B. Let $Z = \mathbb{1}_{\{R_{\mu_A}(h) - \inf_{f\in\tilde{\mathcal{F}}} R_{\mu_A}(f) \leq \varepsilon\}}$. If $Z = 1$, then by the above deliberations we conclude $d_H(A, h(B)) \leq \frac{d}{4}$ and thus, given $h(B)$, A ranges over a set of size $\sum_{i=0}^{\frac{d}{4}}\binom{n}{i} \leq 2^{H\left(\frac{1}{4}\right)d}$. Thus, we get (using data processing and the definition of conditional entropy)

$$\begin{aligned} I(A:B) &\geq I(A:h(B)) = H(A) - H(A|h(B)) \\ &\geq H(A) - H(A|h(B),Z) - H(Z) \end{aligned}$$

$$
\begin{aligned}
&= H(A) - \underbrace{\mathbb{P}[Z=1]}_{\leq 1}\underbrace{H(A|h(B),Z=1)}_{\leq H\left(\frac{1}{4}\right)d} \\
&\quad - \underbrace{\mathbb{P}[Z=0]}_{\leq \delta}\underbrace{H(A|h(B),Z=0)}_{\leq d} - \underbrace{H(Z)}_{\leq H(\delta)} \\
&\geq d - H\left(\frac{1}{4}\right)d - \delta d - H(\delta) \\
&= \left(1 - H\left(\frac{1}{4}\right) - \delta\right)d - H(\delta),
\end{aligned}
$$

in particular $I(A:B) \geq \Omega(d)$. (Here we use our assumption on δ.)

Now we show $I(A:B) \leq m \cdot I(A:B_1)$. We reproduce the reasoning provided in Arunachalam and de Wolf (2018) for completeness:

$$
\begin{aligned}
I(A:B) &= S(B) - S(B|A) \\
&= S(B) - \sum_{i=1}^{m} S(B_i|A) \\
&\leq \sum_{i=1}^{m} S(B_i) - S(B_i|A) \\
&= \sum_{i=1}^{m} I(A:B_1).
\end{aligned}
$$

Here, the first step is by definition, the second uses the product structure of the subsystem B, the third follows from subadditivity of the entropy and the last is again by definition.

And finally, we prove an upper bound on $I(A:B_1)$. To this end, we have to study the reduced state

$$
\sigma_{AB_1} = \frac{1}{2^d} \sum_{a\in\{0,1\}^d} |a\rangle\langle a| \otimes \rho_a.
$$

More precisely, we have

$$
I(A:B_1) = S(A) + S(B_1) - S(AB_1),
$$

and thus have to study the entropies of σ_{AB_1} as well as those of the reduced states σ_A and σ_{B_1}. As $A \sim \text{Uniform}\left(\{0,1\}^d\right)$, we have $S(A) = d$. Now we consider the reduced state

$$
\begin{aligned}
\sigma_{B_1} &= \frac{1}{2^d} \sum_{a\in\{0,1\}^d} \rho_a \\
&= \sum_{i=1}^{d} |s_i\rangle\langle s_i| \otimes \left(\left(\frac{1}{2^d}\sum_{a\in\{0,1\}^d} \mu_a(i,0)\right)|\psi_0\rangle\langle\psi_0| \right. \\
&\quad \left. + \left(\frac{1}{2^d}\sum_{a\in\{0,1\}^d} \mu_a(i,1)\right)|\psi_1\rangle\langle\psi_1|\right).
\end{aligned}
$$

Here, we have

$$
\frac{1}{2^d}\sum_{a\in\{0,1\}^d} \mu_a(i,0) = \frac{1}{2d} = \frac{1}{2^d}\sum_{a\in\{0,1\}^d} \mu_a(i,1).
$$

By Lemma 4 we know that $\frac{1}{2d}|\psi_0\rangle\langle\psi_0| + \frac{1}{2d}|\psi_1\rangle\langle\psi_1|$ has non-zero eigenvalues $\mu_{1/2} = \frac{1}{2d}(1 \pm |\langle\psi_0|\psi_1\rangle|)$ and due to the block-diagonal structure of σ_{B_1} we conclude that the non-zero eigenvalues of σ_{B_1} are also $\mu_{1/2}$, each of multiplicity d. In particular, we have

$$
\begin{aligned}
S(\sigma_{B_1}) &= d \cdot (-\mu_1 \log(\mu_1) - \lambda_2 \log(\mu_2)) \\
&= \log(2d) - \frac{1}{2}\left(\log(1 - |\langle\psi_0|\psi_1\rangle|^2) \right. \\
&\quad \left. + |\langle\psi_0|\psi_1\rangle| \log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right)\right).
\end{aligned}
$$

Similarly, we see that the non-zero eigenvalues of σ_{AB_1} are

$$
\frac{1}{2^d}\lambda_{1/2} = \frac{1}{2^d}\cdot\frac{1}{2d}\left(1 \pm |\langle\psi_0|\psi_1\rangle|\sqrt{1 + \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}\right),
$$

each of multiplicity $d \cdot 2^d$ and that therefore

$$
\begin{aligned}
S(\sigma_{AB_1}) &= d + \log(2d) - \frac{1}{2}\left(\log\left(1 - |\langle\psi_0|\psi_1\rangle|^2\left(1 + \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}\right)\right)\right. \\
&\quad \left. + |\langle\psi_0|\psi_1\rangle|\sqrt{1 + \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}}{1-|\langle\psi_0|\psi_1\rangle|\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}}\right)\right).
\end{aligned}
$$

If we combine these expressions for the different entropies, we obtain

$$
\begin{aligned}
I(A:B_1) &= S(A) + S(B_1) - S(AB_1) \\
&= \frac{1}{2}\left(\log\left(1 - |\langle\psi_0|\psi_1\rangle|^2 - \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}(1-|\langle\psi_0|\psi_1\rangle|^2)\right) - \log\left(1-|\langle\psi_0|\psi_1\rangle|^2\right)\right) \\
&\quad + \frac{|\langle\psi_0|\psi_1\rangle|}{2}\left(\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}}{1-|\langle\psi_0|\psi_1\rangle|\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}}\right)\right. \\
&\quad \left. - \log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right)\right).
\end{aligned}
$$

We now use Taylor's theorem to understand the scaling of the different terms with ε. First, we have (by Taylor-expanding $\log(1-|\langle\psi_0|\psi_1\rangle|^2 - x)$ around $x=0$)

$$
\begin{aligned}
&\log\left(1 - |\langle\psi_0|\psi_1\rangle|^2 - \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}(1-|\langle\psi_0|\psi_1\rangle|^2)\right) \\
&\quad - \log\left(1-|\langle\psi_0|\psi_1\rangle|^2\right) \\
&= \frac{1}{1-|\langle\psi_0|\psi_1\rangle|^2}\cdot\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}(1-|\langle\psi_0|\psi_1\rangle|^2) + \mathcal{O}(\varepsilon^4) \\
&= -\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2} + \mathcal{O}(\varepsilon^4).
\end{aligned}
$$

Moreover, using the Taylor expansions

$$
\log\left(\frac{1+a\sqrt{1+x}}{1-a\sqrt{1+x}}\right) = \log\left(\frac{1+a}{1-a}\right) + \frac{ax}{1-a^2} + \mathcal{O}(x^2)
$$

around $x=0$ (with $a>0$) and

$$
\begin{aligned}
&\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}} \\
&= 1 + \frac{1}{2}\cdot\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2} + \mathcal{O}(\varepsilon^4)
\end{aligned}
$$

we now obtain

$$
\begin{aligned}
&\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}}{1-|\langle\psi_0|\psi_1\rangle|\sqrt{1+\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2}}}\right) \\
&\quad - \log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right) \\
&= \left(1+\frac{1}{2}\cdot\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2} + \mathcal{O}(\varepsilon^4)\right) \\
&\quad \cdot\left(\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right) + \frac{|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|^2}\cdot\frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{|\langle\psi_0|\psi_1\rangle|^2} + \mathcal{O}(\varepsilon^4)\right) \\
&\quad - \log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right) \\
&= \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\left(\frac{1}{|\langle\psi_0|\psi_1\rangle|} + \frac{1-|\langle\psi_0|\psi_1\rangle|^2}{2|\langle\psi_0|\psi_1\rangle|}\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right)\right) + \mathcal{O}(\varepsilon^4).
\end{aligned}
$$

Plugging these approximations back in gives us

$$
\begin{aligned}
I(A:B_1) &= \frac{64\varepsilon^2}{\|\sigma_0-\sigma_1\|_1^2}\cdot\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{4|\langle\psi_0|\psi_1\rangle|} \\
&\quad \times\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right) + \mathcal{O}(\varepsilon^4) = \mathcal{O}(\varepsilon^2).
\end{aligned}
$$

Now combining our mutual information lower and upper bounds yields

$$
\Omega(d) \le I(A:B) \le m\cdot I(A:B_1) \le m\cdot\mathcal{O}(\varepsilon^2),
$$

which after rearranging becomes

$$
m \ge \Omega\left(\frac{d}{\varepsilon^2}\right),
$$

as desired. □

Detailed Proof of Lemma 5 As $\mathscr{F}$ is non-trivial, there exist $f_1, f_2 \in \mathscr{F}$ and $x_1, x_2 \in \mathscr{X}$ s.t. $f_1(x_1) = f_2(x_1) = \sigma_0$ and $f_1(x_2) = \sigma_0 \neq \sigma_1 = f_2(x_2)$. Now consider the distribution μ on $\mathscr{X}$ defined by

$$\mu(x_1) = 1 - \lambda, \quad \mu(x_2) = \lambda,$$

where $\lambda \in (0, 1)$ is to be chosen later in the proof.

The risk of a hypothesis $h \in \mathscr{D}^{\mathscr{X}}$ w.r.t. μ if the target concept is f_i is given by

$$R_{\mu, f_i}(h) = \frac{1-\lambda}{2} \|h(x_1) - f_i(x_1)\|_1 + \frac{\lambda}{2} \|h(x_2) - f_i(x_2)\|_1 ,$$

so in particular we have

$$R_{\mu, f_i}(f_j) = \begin{cases} 0 & \text{if } i = j \\ \frac{\lambda}{2} \|\sigma_0 - \sigma_1\|_1 & \text{if } i \neq j \end{cases}.$$

So if we choose $\lambda = \frac{2\varepsilon}{\|\sigma_0 - \sigma_1\|_1} < 1$, then the learning requirement for $\mathscr{A}$ implies that with probability $\geq 1 - \delta$, $\mathscr{A}$ correctly identifies whether the target concept is f_1 or f_2. As the algorithm has access to the underlying distribution only via the training data, this means that the algorithm has to be able to distinguish the corresponding training data ensembles with probability $\geq 1 - \delta$. Here, we observe that the training data being drawn i.i.d. according to $\mu_{\pm}$ is equivalent to the learning algorithm having access to m copies of the state

$$\rho_i = (1-\lambda)|x_1\rangle\langle x_1| \otimes \sigma_0 + \lambda |x_2\rangle\langle x_2| \otimes f_i(x_2), \quad i = 1, 2.$$

The optimal success probability for distinguishing between two quantum states is a well-studied object in quantum information theory. It can be characterized by the trace distance between the two states and is given (in our case) by (see Nielsen and Chuang 2009)

$$p_{\text{opt}} = \frac{1}{2}(1 + \frac{1}{2} \left\| \rho_1^{\otimes m} - \rho_2^{\otimes m} \right\|_1).$$

As the trace distance of tensor products is not that easy to deal with, we will instead work with the fidelity defined as $F(\rho, \sigma) := \text{tr}[\sqrt{\rho^{\frac{1}{2}} \sigma \rho^{\frac{1}{2}}}]$. According to the Fuchs-van de Graaf inequalities (see Nielsen and Chuang 2009, Section 9.2.3) we have

$$\begin{aligned} &\frac{1}{2} \left\| \rho_1^{\otimes m} - \rho_2^{\otimes m} \right\|_1 \leq \sqrt{1 - F(\rho_1^{\otimes m}, \rho_2^{\otimes m})^2} \\ &= \sqrt{1 - F(\rho_1, \rho_2)^{2m}}, \end{aligned}$$

where the last steps uses multiplicativity of the fidelity under tensor products. Now we require $p_{\text{opt}} \geq 1 - \delta$ and rearrange to obtain

$$F(\rho_1, \rho_2)^{2m} \leq 4\delta(1-\delta)$$

or equivalently after taking logarithms

$$m \geq \frac{\log(4\delta(1-\delta))}{\log(F(\rho_1, \rho_2)^2)}.$$

Now we use again the Fuchs-van de Graaf inequalities which tell us (after rearranging)

$$1 - \frac{1}{2} \|\rho_1 - \rho_2\|_1 \leq F(\rho_1, \rho_2) \leq \sqrt{1 - \frac{1}{4} \|\rho_1 - \rho_2\|_1^2}$$

to obtain that

$$\begin{aligned} m &\geq \frac{\log(4\delta(1-\delta))}{\log(F(\rho_1, \rho_2)^2)} = \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\log\left(\frac{1}{F(\rho_1, \rho_2)^2}\right)} \\ &\geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\log\left(\frac{1}{(1 - \frac{1}{2}\|\rho_1 - \rho_2\|_1)^2}\right)} \geq \frac{\log(4\delta(1-\delta))}{2\log(1 - \frac{1}{2}\|\rho_1 - \rho_2\|_1)}. \end{aligned}$$

It is easy to see that $\|\rho_1 - \rho_2\|_1 = \lambda \|\sigma_0 - \sigma_1\|_1 = 2\varepsilon$. Now Taylor expansion of the logarithm gives

$$m \geq \Omega\left(\frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon}\right),$$

as desired. □

Detailed Proof of Theorem 4 Let $S = (s_0, \ldots, s_d) \in \mathscr{X}$ be a set shattered by $\tilde{\mathscr{F}}$, define

$$\mu(s_0) = 1 - \lambda, \quad \mu(s_i) = \frac{\lambda}{d} \quad \forall 1 \leq i \leq d,$$

with $\lambda \in (0, 1)$ to be chosen later. By shattering, $\forall a \in \{0, 1\}^d \; \exists f_a \in \tilde{\mathscr{F}}$ s.t.

$$f_a(s_0) = 0 \quad \text{and} \quad f_a(s_i) = a_i \quad \forall 1 \leq i \leq d.$$

Observe that w.r.t. a distribution μ and target concept f_a, another concept f_b has error

$$d_H(a, b) \cdot \frac{\lambda}{d} \cdot \frac{\|\sigma_0 - \sigma_1\|_1}{2}.$$

So if we pick $\lambda = \frac{8\varepsilon}{\|\sigma_0 - \sigma_1\|_1}$, then by the learning requirement, with probability $\geq 1 - \delta$, $\mathscr{A}$ has to output a hypothesis h that when evaluated on S yields a label vector that is $\frac{d}{4}$-close to the true underlying string in Hamming distance.

Denote by $A \sim \text{Uniform}\left(\{0, 1\}^d\right)$ a random variable describing the unknown underlying string, let $B = B_1 \ldots B_m$ be the corresponding quantum training data system. We want to repeat the three-step reasoning from the proof of Theorem 3. The first two steps work exactly as before. Step 3 will be slightly different. Again we have

$$I(A : B_1) = S(A) + S(B_1) - S(AB_1), \quad \text{and} \quad S(A) = d.$$

In this case, the relevant composite state is

$$\sigma_{AB_1} = \frac{1}{2^d} \sum_{a\in\{0,1\}^d} |a\rangle\langle a| \otimes \rho_a,$$

where $\rho_a = \sum_{j=0}^{d} \mu(s_j)|s_j\rangle\langle s_j| \otimes f_a(s_j) = (1-\lambda)|s_0\rangle\langle s_0| \otimes \sigma_0 + \frac{\lambda}{d} \sum_{j=1}^{d} |s_j\rangle\langle s_j| \otimes \sigma_{a_j}$.

We now again use Lemma 4 to compute eigenvalues and thus entropies. (Here our assumption that σ_0 and σ_1 are pure enters the proof.) We obtain

- Each ρ_a has non-zero eigenvalues $1-\lambda$ of multiplicity 1 and $\frac{\lambda}{d}$ of multiplicity d.
- $\sigma_{B_1} = \frac{1}{2^d} \sum_{a\in\{0,1\}^d} \left((1-\lambda)|s_0\rangle\langle s_0| \otimes \sigma_0 + \frac{\lambda}{d} \sum_{j=1}^{d} |s_j\rangle\langle s_j| \otimes \sigma_{a_j}\right) = (1-\lambda)|s_0\rangle\langle s_0| \otimes \sigma_0 + \frac{\lambda}{d} \sum_{j=1}^{d} |s_j\rangle\langle s_j| \otimes \left(\frac{1}{2}\sigma_0 + \frac{1}{2}\sigma_1\right)$ has non-zero eigenvalues $1-\lambda$ of multiplicity 1 and $\frac{\lambda}{d}\lambda_{1/2}$ of multiplicity d, where $\lambda_{1/2} = \frac{1\pm|\langle\psi_0|\psi_1\rangle|}{2}$.
- σ_{AB_1} has non-zero eigenvalues $\frac{1}{2^d}(1-\lambda)$ of multiplicity 2^d and $\frac{\lambda}{d\cdot 2^d}$ of multiplicity $d\cdot 2^d$.

With this we can now compute the relevant entropies and obtain

$$\begin{aligned} S(B_1) &= S(\sigma_{B_1}) \\ &= -(1-\lambda)\log(1-\lambda) + d\left(-\frac{\lambda}{d}\lambda_1 \log\left(\frac{\lambda}{d}\lambda_1\right) - \frac{\lambda}{d}\lambda_2 \log\left(\frac{\lambda}{d}\lambda_2\right)\right) \\ &= -(1-\lambda)\log(1-\lambda) - \lambda\left(\lambda_1 \log\left(\frac{\lambda}{d}\lambda_1\right) + \lambda_2 \log\left(\frac{\lambda}{d}\lambda_2\right)\right), \end{aligned}$$

as well as

$$\begin{aligned} S(AB_1) &= S(\sigma_{AB_1}) \\ &= 2^d\left(-\frac{1}{2^d}(1-\lambda)\log\left(\frac{1}{2^d}(1-\lambda)\right) - d\cdot\frac{\lambda}{d\cdot 2^d}\log\left(\frac{\lambda}{d\cdot 2^d}\right)\right) \\ &= -(1-\lambda)\log\left(\frac{1-\lambda}{2^d}\right) - \lambda\log\left(\frac{\lambda}{d\cdot 2^d}\right). \end{aligned}$$

Hence, we now have

$$\begin{aligned} I(A:B_1) &= S(A) + S(B_1) - S(AB_1) \\ &= -\frac{\lambda}{2}\underbrace{\left(\log\left(\frac{1-|\langle\psi_0|\psi_1\rangle|^2}{4}\right) + |\langle\psi_0|\psi_1\rangle|\log\left(\frac{1+|\langle\psi_0|\psi_1\rangle|}{1-|\langle\psi_0|\psi_1\rangle|}\right)\right)}_{\leq 0 \text{ because } |\langle\psi_0|\psi_1\rangle|\in[0,1]} \\ &= \mathcal{O}(\varepsilon). \end{aligned}$$

Now we can finish the proof by combining steps 1, 2 and 3 as before. □

Appendix 2. A physical motivation for our notion of risk

In our definition of the risk R_μ we use the trace distance. As the latter is a well-established measure of distinguishability of quantum states, it presents itself as a natural candidate loss function. Here, we give a more explicit operational reasoning as to why we choose to use the trace distance.

Imagine the learning task as a competition between two parties, a learner and a teacher. We assume that both parties obey the laws of quantum physics. The teacher knows (a classical description of) the probability distribution $\mu \in \mathrm{Prob}(\mathscr{X}\times\mathscr{D})$ and will provide corresponding training data to the learner during a training phase. The learner's goal is to persuade the teacher in a test phase that she has managed to learn the distribution μ, which was unknown to her in advance, i.e., that she has produced a good hypothesis $h : \mathscr{X} \to \mathscr{D}$.

We first give an informal description of the test phase: The teacher prepares another (independent) example (x, ρ) drawn from μ. She then sends x to the learner. The latter applies her hypothesis h to prepare the quantum state $h(x)$ which she then sends back to the teacher. The teacher now uses this one copy of $h(x)$ and her knowledge of μ to evaluate whether the learner made a good prediction. As also the teacher is restricted by quantum theory, she can only do so by performing a measurement.

We now discuss the choice of measurement of the teacher in more detail. On the one hand, the teacher wants to maximize the probability of detecting a wrong prediction. On the other hand, she does not want to be unfair, so at the same time she tries to maximize the probability of detecting

a correct prediction. In summary, the teacher wants to choose a 2-outcome measurement $\{E_{accept}, E_{reject}\}$ that maximizes

$$\mathrm{tr}[E_{accept}\sigma_i] + \mathrm{tr}[E_{reject}\sigma_j],$$

where $\sigma_i = \rho$ and $\sigma_j \in \mathscr{D} \setminus \{\rho\}$. As she knows (a classical description of) the state $\rho \in \mathscr{D}$ and that $h(x) \in \mathscr{D}$, she can achieve this by picking $\{E_{accept}, E_{reject}\}$ to be the optimal measurement for minimum error discrimination of $\mathscr{D}$ (where the states are taken with equal prior probabilities (see Watrous 2018, Theorem 3.4)). The measurement is basically the same independently of whether $\rho = \sigma_1$ or $\rho = \sigma_2$, only the outcome labels are interchanged.

Now the expected probability of the trainer rejecting the learner's prediction is

$$\int_{X\times\mathscr{D}} \mathrm{tr}[E_{reject}(\rho)h(x)]\,\mathrm{d}\mu(x,\rho).$$

The optimal measurement satisfies

$$\mathrm{tr}[E_{accept}\sigma_i] + \mathrm{tr}[E_{reject}\sigma_j] = \frac{1}{2}\left(1 + \frac{1}{2}\|\sigma_0 - \sigma_1\|_1\right).$$

It is easy to see that under the additional assumption that σ_0 and σ_1 have the same purity, i.e., $\mathrm{tr}[\sigma_0^2] = \mathrm{tr}[\sigma_1^2]$, the rejection probabilities are symmetric, namely

$$\mathrm{tr}[E_{accept}\sigma_j] = \mathrm{tr}[E_{reject}\sigma_i] = \frac{1}{2}\left(1 - \frac{1}{2}\|\sigma_0 - \sigma_1\|_1\right)$$

and similarly

$$\mathrm{tr}[E_{accept}\sigma_i] = \mathrm{tr}[E_{reject}\sigma_j] = \frac{1}{2}\left(1 + \frac{1}{2}\|\sigma_0 - \sigma_1\|_1\right).$$

With this we now obtain when comparing the achieved with the optimal expected rejection probability

$$\begin{aligned}
&\int_{X\times\mathscr{D}} \mathrm{tr}[E_{reject}(\rho)h(x)]\,\mathrm{d}\mu(x,\rho) \\
&\quad - \inf_{g:\mathscr{X}\to\mathscr{D}} \int_{X\times\mathscr{D}} \mathrm{tr}[E_{reject}(\rho)g(x)]\,\mathrm{d}\mu(x,\rho) \\
&= \int_{X\times\mathscr{D}} \frac{\|\rho - h(x)\|_1}{4}\,\mathrm{d}\mu(x,\rho) = \frac{1}{2}R_\mu(h).
\end{aligned}$$

So we have recovered our notion of risk, at least in the case of states of equal purity, from a more basic analysis of the test phase.

Note that a similar analysis could be performed also in the case of more than two quantum labels. There, the teacher's measurement would be the optimal measurement for minimum error discrimination of ρ and $\frac{1}{|\mathscr{D}|-1}\sum_{\sigma\in\mathscr{D}\setminus\{\rho\}}\sigma$. Unfortunately, no closed-form expressions for the corresponding success probabilities are known. We do, however, see that in this scenario, using the trace distance as loss function would be too pessimistic from the perspective of the learner. As the teacher does not know the prediction state prepared by the learner, the teacher has to solve a state discrimination problem taking into account all possible label states.

Appendix 3. The Holevo-Helstrom strategy

The naive learning strategy based on the Holevo-Helstrom measurement is the following:

Holevo-Helstrom strategy

Given: Quantum training data $S = \{(x_i, \rho_i)\}_{i=1}^m$
Output: Hypothesis $\hat{h} : \mathscr{X} \to \mathscr{D}$
Algorithm:

1. For each i: Perform a Holevo-Helstrom measurement on ρ_i. Let
$$y_i = \begin{cases} 1 & \text{if } E_1 \text{ is accepted} \\ 0 & \text{if } E_1 \text{ is rejected} \end{cases}.$$
2. Let $\tilde{S} = \{(x_i, y_i)\}_{i=1}^m \in (\mathscr{X}\times\{0,1\})^m$. Then one can view (x_i, y_i) as being drawn independently according to the probability measure ν on $\mathscr{X}\times\{0,1\}$ which has
$$\nu_1(x) = \mu_1(x) = \mu(x,\sigma_0) + \mu(x,\sigma_1)$$
as the first marginal and
$$\begin{aligned}\nu(y|x) = {} & \delta_{y0}\left(\mu(\sigma_1|x)\mathrm{tr}[\sigma_1 E_0] + \mu(\sigma_0|x)\mathrm{tr}[\sigma_0 E_0]\right) \\ & + \delta_{y1}\left(\mu(\sigma_1|x)\mathrm{tr}[\sigma_1 E_1] + \mu(\sigma_0|x)\mathrm{tr}[\sigma_0 E_1]\right).\end{aligned}$$
as the conditional probability distribution of y given x.
3. Use a classical learning algorithm for binary classification to find $\{\hat{g}\} \in \tilde{\mathscr{F}} := \{\tilde{f} : \mathscr{X} \to \{0,1\} \mid \exists f \in \mathscr{F} : f(x) = \sigma_{\tilde{f}(x)}\ \forall x \in \mathscr{X}\}$ s.t. $\tilde{R}_\nu(\hat{g}) := \mathbb{P}_{(x,y)\sim\nu}[y \neq \{\hat{g}\}(x)]$ is small.
4. Define $\{\hat{h}\} : \mathscr{X} \to \mathscr{D}$ via $\{\hat{h}\}(x) = \sigma_{\{\hat{g}\}(x)}$ and output $\{\hat{h}\}$ as hypothesis.

The remainder of this section is devoted to studying the performance of this simple learning procedure. Note that we leave open for now the classical learning algorithm to be used, we first work towards characterizing the true risk $R_\mu(h)$ in terms of the intermediate classical risk $\tilde{R}_\nu(g)$.

In the following we will often make use of the fact that when identifying $i \leftrightarrow \sigma_i$, the probability measure μ on $\mathscr{X}\times\mathscr{D}$ gives rise to a probability measure on $\mathscr{X}\times\{0,1\}$. We will abuse notation and also denote the latter measure

by μ, however, which measure is meant will always be clear from the context.

Recall that $R_\mu(h) = \frac{\|\sigma_0-\sigma_1\|}{2}\mathbb{P}_{(x,\rho)\sim\mu}[h(x) \neq \rho]$. We now derive a similar expression for $\tilde{R}_\nu(g)$.

Lemma C.1 *With the notation as in the Holevo-Helstrom strategy (in particular $h(x) = \sigma_{g(x)}$) it holds that*

$$\tilde{R}_\nu(g) = \frac{\|\sigma_0-\sigma_1\|_1}{2}\mathbb{P}_{(x,\rho)\sim\mu}[h(x) \neq \rho] + \mathrm{tr}[\sigma_0 E_1] + (\mathrm{tr}[\sigma_1 E_0] - \mathrm{tr}[\sigma_0 E_1])\mathbb{E}_{\mu_1}[g].$$

Proof This can be shown by direct computation using the definition of ν:

$$\begin{aligned}\tilde{R}_\nu(g) &= \int_{\mathscr{X}\times\{0,1\}} |y-g(x)|\mathrm{d}\nu(x,y)\\ &= \int_{\mathscr{X}}\left(\int_{\{0,1\}} |y-g(x)|\mathrm{d}\nu(y|x)\right)\mathrm{d}\nu_1(x)\\ &= \int_{\mathscr{X}} (|1-g(x)|(\mu(\sigma_1|x)\mathrm{tr}[\sigma_1 E_1] + \mu(\sigma_0|x)\mathrm{tr}[\sigma_0 E_1])\\ &\quad + |g(x)|(\mu(\sigma_1|x)\mathrm{tr}[\sigma_1 E_0] + \mu(\sigma_0|x)\mathrm{tr}[\sigma_0 E_0]))\,\mathrm{d}\mu_1(x)\end{aligned}$$

Now we use the specific property of the Holevo-Helstrom measurement that $\mathrm{tr}[(\sigma_1-\sigma_0)E_1] = \frac{\|\sigma_0-\sigma_1\|}{2}$. Moreover, as $g(x) \in \{0,1\}$, we have $|1-g(x)| = 1-g(x)$ and $|g(x)| = g(x)$. Thus, we obtain

$$\begin{aligned}\tilde{R}_\nu(g) &= \frac{\|\sigma_0-\sigma_1\|}{2}\int_{\mathscr{X}} ((1-g(x))\mu(\sigma_1|x) + g(x)\mu(\sigma_0|x))\,\mathrm{d}\mu_1(x)\\ &\quad + \int_{\mathscr{X}} ((1-g(x))\mathrm{tr}[\sigma_0 E_1] + g(x)\mathrm{tr}[\sigma_1 E_0])\,\mathrm{d}\mu_1(x)\\ &= \frac{\|\sigma_0-\sigma_1\|}{2}\mathbb{P}_{(x,\rho)\sim\mu}[h(x) \neq \rho] + \mathrm{tr}[\sigma_0 E_1]\\ &\quad + (\mathrm{tr}[\sigma_1 E_0] - \mathrm{tr}[\sigma_0 E_1])\mathbb{E}_{\mu_1}[g],\end{aligned}$$

where the last step uses $h(x) = \sigma_{g(x)}$. □

This allows us to easily compare the true and the intermediate risk and obtain

$$\tilde{R}_\nu(g) - R_\mu(h) = \mathrm{tr}[\sigma_0 E_1](1-2\mathbb{E}_{\mu_1}[g]) + \left(1-\frac{\|\sigma_0-\sigma_1\|}{2}\right)\mathbb{E}_{\mu_1}[g].$$

As $g(x) \in \{0,1\}\ \forall x \in \mathscr{X}$ and in particular $0 \leq \mathbb{E}_{\mu_1}[g] \leq 1$, this gives rise to the following

Corollary 2 *With the notation as in the Holevo-Helstrom strategy it holds that*

$$\tilde{R}_\nu(g) - \max\{\mathrm{tr}[\sigma_0 E_1], \mathrm{tr}[\sigma_1 E_0]\} \leq R_\mu(h) \leq \tilde{R}_\nu(g) - \min\{\mathrm{tr}[\sigma_0 E_1], \mathrm{tr}[\sigma_1 E_0]\}.$$

We can extend this to a comparison between the excess risks

$$R_\mu(h) - R^*_{\mu,\mathscr{F}} := R_\mu(h) - \inf_{\eta\in\mathscr{F}} R_\mu(\eta) \text{ and } \tilde{R}_\nu(g) - \tilde{R}^*_{\nu,\tilde{\mathscr{F}}} := \tilde{R}_\nu(g) - \inf_{\gamma\in\tilde{\mathscr{F}}} \tilde{R}_\nu(\gamma)$$

which are the quantities of interest for agnostic learning scenarios.

Corollary 3 *With the notation as in the Holevo-Helstrom strategy it holds that*

$$\tilde{R}_\nu(g) - \tilde{R}^*_{\nu,\tilde{\mathscr{F}}} - |\mathrm{tr}[\sigma_0 E_1] - \mathrm{tr}[\sigma_1 E_0]| \leq R_\mu(h) - R^*_{\mu,\mathscr{F}} \leq \tilde{R}_\nu(g) - \tilde{R}^*_{\nu,\tilde{\mathscr{F}}} + |\mathrm{tr}[\sigma_0 E_1] - \mathrm{tr}[\sigma_1 E_0]|$$

So we see that solving the classical learning task in step 3 of the Holevo-Helstrom strategy does not necessarily imply success at the overall learning task if the target accuracy is $\varepsilon < |\mathrm{tr}[\sigma_0 E_1] - \mathrm{tr}[\sigma_1 E_0]|$. This problem is addressed by the noise-corrected Holevo-Helstrom strategy presented in Section 4.

Remark 4 We want to shortly discuss a special case in which the connection between $R_\mu(h)$ and $\tilde{R}_\nu(g)$ takes a particularly appealing form. Namely, assume that σ_0 and σ_1 are such that the corresponding Holevo-Helstrom measurement produces equal probabilities of error, i.e., $\mathrm{tr}[E_0\sigma_1] = \mathrm{tr}[E_1\sigma_0]$. This is clearly not true in general, take, e.g., $\sigma_0 = |0\rangle\langle 0|$ and $\sigma_1 = \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|)$. It does, however, hold true in certain special cases, e.g., if both σ_0 and σ_1 are pure or if σ_0 and σ_1 have the same (non-trivial) purity and $\mathrm{tr}[E_0] = \mathrm{tr}[E_1]$. (The latter is, e.g., satisfied if σ_0 and σ_1 are qubit states of the same (non-zero) purity.)

In this simple case our previous discussion yields $R_\mu(h) = \tilde{R}_\nu(g)$, in particular, if we succeed at the classical binary classification task in step 3, then we also succeed at the overall classification task with quantum labels, so the quantum learning task is reduced to a classical learning problem.

Appendix 4. Sample complexity of binary classification with two-sided classification noise

Here, we discuss the sample complexity of the PAC learning task of binary classification in the presence of (two-sided) classification noise in the realizable scenario. To be in congruence with the literature on this and related problems, we will use a slightly different notation than in the main body of the paper. Namely, we will consider classical input

space $\mathscr{X}$ and classical target space $\{0,1\}$, a concept class $\mathscr{F} \subset \{0,1\}^{\mathscr{X}}$, a probability measure $\mu \in \text{Prob}(\mathscr{X})$, and noise probabilities $0 \leq \eta_0, \eta_1 < \frac{1}{2}$, with which labels are flipped. Moreover, we will work with the 0-1-loss function and denote the corresponding risk of a hypothesis h w.r.t. a target concept f by $\text{err}_\mu(h; f) = \mu[h(x) \neq f(x)]$. Finally, any training data sample S splits the concept class $\mathscr{F}$ into so-called S-equivalence classes, where $f_1, fc \in \mathscr{F}$ are equivalent if and only if $f_1(x) = f_2(x)\ \forall x \in \mathscr{X}$ s.t. $\exists y \in \{0,1\}$ with $(x,y) \in S$.

The basic learning strategy underlying our discussion is Algorithm 1. It is the natural analog of searching for a consistent function in the case of noisy labels. Namely, as such a consistent function will in general not exist, it searches for a function that disagrees with the training data on as few examples as possible.

Algorithm 1 Minimum Disagreement Strategy L (Laird 1988, Algorithm 5.6).

Input: confidence and accuracy parameters $0 < \delta, \varepsilon \leq \frac{1}{2}$, a noise threshold $0 \leq \eta_0, \eta_1 \leq \eta_b < \frac{1}{2}$ and noisy training data $S = \{(x_i, y_i)\}_{i=1}^m$ created from $\mu \in \text{Prob}(\mathscr{X})$ and some $f \in \mathscr{F}$ where 0-labels are corrupted with prob. η_0 and 1-labels are corrupted with prob. η_1, where

$$m \geq \underbrace{\max\{\frac{8}{\varepsilon}\log\left(\frac{6}{\delta}\right), \frac{16d}{\varepsilon}\log\left(\frac{16d}{\varepsilon}\right)\}}_{=:m_1} + \underbrace{\frac{2}{\varepsilon(1-\exp(-\frac{1}{2}(1-2\eta_b)^2))}\ln\left(\frac{1}{d}\left(\max\{\frac{8}{\varepsilon}\log\left(\frac{6}{\delta}\right), \frac{16d}{\varepsilon}\log\left(\frac{16d}{\varepsilon}\right)\}^d + 1\right)\right)}_{=:m_2}.$$

Output: a hypothesis $h \in \mathscr{F}$.

1: Let S_1 consist of the first m_1 examples in S. Let $S_2 = S \setminus S_1$.
2: Select $\mathscr{F}_1 = \{f_1, \ldots, f_N\}$ as representatives of the S_1-equivalence classes induced by S_1, where $N \leq (m_1)^d + 1$.
3: Output a hypothesis in $\mathscr{F}_1$ which minimizes the number of disagreements with S_2.

Theorem 4.1 (see Laird 1988, Theorems 5.7 and 5.33)
The output hypothesis h of Algorithm 1 satisfies $err_\mu(h; f) \leq \varepsilon$.

Laird's original proof that this algorithm solves the PAC learning problem is for the case $\eta_0 = \eta_1$. It is, however, easily generalized to our case because we still assume the same noise bound on both error rates. (We only have to adapt the expression for the error rate and the corresponding Hoeffding bounds.)

In order to apply the reasoning by Hanneke (2016) we need to slightly reformulate the result of this algorithm s.t. we obtain a bound on the error in terms of the sample size. When following the proof of Theorem 5.7 in Laird (1988) we see that m_1 is used to ensure that there is a hypothesis which performs better than some given error threshold and m_2 is used to ensure that such a hypothesis is actually chosen. In particular, if we use the error bound by Blumer et al. (1989) in terms of the sample size, we see that m_2 depends on m_1 as follows:

$$m_2 = \frac{2}{1-\exp(-\frac{1}{2}(1-2\eta_b)^2)} \cdot \frac{m_1}{2} \cdot \frac{1}{d\log\left(\frac{2em_1}{d}\right) + \log\left(\frac{2}{\delta}\right)} \cdot \ln\left(\frac{1}{\delta}(m_1^d + 1)\right).$$

Remark 5 Note that we cannot directly use the tighter error bound in terms of the sample complexity proved by Hanneke (2016) here because Laird's proof explicitly makes use of the strategy employed by Blumer et al. (1989) which works via consistency with a given training sample.

We can now easily bound

$$m = m_1 + m_2 \leq m_1 \cdot \left(1 + \frac{1}{1-\exp(-\frac{1}{2}(1-2\eta_b)^2)} \cdot \frac{1}{\log(e)} \cdot \frac{1}{1 - \frac{d\log\left(\frac{d}{2e}\right)}{d\log(m_1) + \log\left(\frac{2}{\delta}\right)}}\right).$$

If we now further assume that $\delta > 0$ is chosen s.t. $\log\left(\frac{2}{\delta}\right) > 2d\log\left(\frac{d}{2e}\right)$, then we can continue upper bounding this and obtain

$$m = m_1 + m_2 \leq (1 + C(\eta_b))m_1,$$

where we defined $C(\eta_b) := \frac{2}{1-\exp(-\frac{1}{2}(1-2\eta_b)^2)}$. It is easy to check that for $0 \leq \eta_b < \frac{1}{2}$, $C(\eta_b) \leq \frac{4}{(1-2\eta_b)^2}$, which well be used later on.

Hence, using a sample of size $m \geq 2(1 + C(\eta_b))$ for the minimum disagreement strategy with $m_2 = \lceil \frac{C(\eta_b)}{1+C(\eta_b)} m \rceil$ and $m_1 = m - m_2$ gives - using $\frac{m}{2(1+C(\eta_b))} \leq m_1 \leq \frac{m}{1+C(\eta_b)} \leq \frac{m_2}{C(\eta_b)}$—an error guarantee of

$$\begin{aligned} \mathrm{err}_\mu(h; f^*) &\leq \frac{4}{m_1}\left(d \log\left(\frac{2em_1}{d}\right) + \log\left(\frac{2}{\delta}\right)\right) \qquad (4.1) \\ &\leq \frac{8 \cdot (1 + C(\eta_b))}{m}\left(d \log\left(\frac{2em}{d \cdot (1 + C(\eta_b))}\right)\right. \\ &\quad \left. + \log\left(\frac{2}{\delta}\right)\right). \qquad (4.2) \end{aligned}$$

With this suboptimal base learner we will now follow the strategy by Hanneke (2016) in order to build a better learner from it. Note that Hanneke's proof includes several steps in which the existence of a function consistent with the respective subsample is ensured. This is not necessary in our case because the minimum disagreement strategy does not require a consistent function to exist.

We recall the algorithm for preprocessing the training data to generate subsamples as introduced in Hanneke (2016) in our Algorithm 2.

Algorithm 2 Subsample Generation Algorithm $\mathbb{A}(\cdot, \cdot)$ Hanneke (2016).

Input: two finite sets S and T.
Output: a finite set $\mathbb{A}(S; T)$ of subsets of $S \cup T$.

1: **if** $|S| \leq 3$, **then**
2: Output $\{S \cup T\}$.
3: **else**
4: Divide $S = \{s_1, \ldots, s_{|S|}\}$ into subsets in the following way:

$$\begin{aligned} S_0 &= \{s_1, \ldots, s_{|S|-3\lfloor |S|/4 \rfloor}\}, \\ S_1 &= \{s_{|S|-3\lfloor |S|/4 \rfloor+1}, \ldots, s_{|S|-2\lfloor |S|/4 \rfloor}\}, \\ S_2 &= \{s_{|S|-2\lfloor |S|/4 \rfloor+1}, \ldots, s_{|S|-\lfloor |S|/4 \rfloor}\}, \\ S_3 &= \{s_{|S|-\lfloor |S|/4 \rfloor+1}, \ldots, s_{|S|}\}. \end{aligned}$$

5: **end if**
6: Return $\mathbb{A}(S_0; S_2 \cup S_3 \cup T) \cup \mathbb{A}(S_0; S_1 \cup S_3 \cup T) \cup \mathbb{A}(S_0; S_1 \cup S_2 \cup T)$.

Theorem 4.2 *Let* $\varepsilon \in (0, 1)$, $\delta \in (0, 2 \cdot (\frac{2e}{d})^d)$ *and* $\eta_b \in (0, \frac{1}{2})$. *Let* $\mathscr{F} \subset \{0, 1\}^{\mathscr{X}}$ *be a function class of VC-dimension* d. *Then* $m = m(\varepsilon, \delta) = \mathscr{O}\left(\frac{1}{\varepsilon(1-2\eta_b)^2}\left(d + \log\left(\frac{1}{\delta}\right)\right)\right)$ *noisy examples from a function in* $\mathscr{F}$ *are sufficient for binary classification in the presence of two-sided classification noise with error probabilities* $0 \leq \eta_0, \eta_1 < \eta_b$ *with accuracy* ε *and confidence* $1 - \delta$.

Proof This proof is analogous to the proof of Theorem 2 in Hanneke (2016) with some minor simplifications and adaptations and is given here only for the sake of completeness.

Fix an $f^* \in \mathscr{F}$ and a probability measure μ over $\mathscr{X}$. Denote by $S = S_{1:m}$ the corresponding noisy training data. For any classifier h denote by $ER(h) = \{x \in \mathscr{X} | h(x) \neq f^*(x)\}$ the set of instances on which h errs.

Fix $c = 7200$. We will show by strong induction that $\forall m' \in \mathbb{N}$, $\forall \delta' \in (0, \ldots)$ and for all finite sequences T' with probability $\geq 1 - \delta'$ the classifier

$$\hat{h}_{m',T'} = \text{Majority}\left(L(\mathbb{A}(S_{1:m'}; T'))\right)$$

satisfies the error bound

$$\mathrm{err}_\mu(\hat{h}_{m',T'}, f^*) \leq \frac{cC(\eta_b)}{1 + m'}\left(d + \ln\left(\frac{18}{\delta'}\right)\right). \qquad (4.3)$$

As base case consider $m' \leq C(\eta_b)c \cdot \ln(18e) - 1$. In this case, for any $\delta' \in (0, 1)$ and for any finite sequence T', we trivially have

$$\begin{aligned} \mathrm{err}_\mu(\hat{h}_{m',T'}, f^*) &\leq 1 \\ &\leq \frac{c \cdot C(\eta_b)}{1 + m'}(d + \ln(18)) \\ &\leq \frac{c \cdot C(\eta_b)}{1 + m'}\left(d + \ln\left(\frac{18}{\delta'}\right)\right), \end{aligned}$$

as desired.

For the induction step, assume that for some $m > C(\eta_b)c \cdot \ln(18e) - 1$ for all $m' \in \mathbb{N}$ with $m' < m$, for all $\delta'(0, 2 \cdot (\frac{2e}{d})^d)$ and for all finite sequences T' with probability $\geq 1 - \delta'$, (4.3) holds.

Note that by our choice of c we have $C(\eta_b)c \cdot \ln(18e) - 1 \geq 3$. Thus, $|S_{1:m}| \geq 4$ and therefore $\mathbb{A}(S_{1:m}; T)$ returns in step 3. Let S_0, S_1, S_2, S_3 be as in $\mathbb{A}(S; T)$. Denote $T_1 = S_2 \cup S_3 \cup T$, $T_2 = S_1 \cup S_3 \cup T$, $T_3 = S_1 \cup S_2 \cup T$ and $h_i = \text{Majority}(L(\mathbb{A}(S_0; T_i)))$ for each $i \in \{1, 2, 3\}$.

Note that $S_0 = S_{1:(m-3\lfloor \frac{m}{4} \rfloor)}$. As $m \geq 4$, $1 \leq m - 3\lfloor \frac{m}{4} \rfloor < m$. Also, $h_i = \hat{h}_{(m-3\lfloor \frac{m}{4} \rfloor), T_i}$. So by the induction hypothesis applied under the conditional distribution given S_1, S_2, S_3, which are independent of S_0, combined with the law of total probability, for every $i \in \{1, 2, 3\}$ there exists an event E_i of probability $\geq 1 - \frac{\delta}{9}$ on which

$$\begin{aligned} \mu[ER(h_i)] &\leq \frac{cC(\eta_b)}{1 + |S_0|}\left(d + \ln\left(\frac{9 \cdot 18}{\delta}\right)\right) \\ &\leq \frac{4cC(\eta_b)}{m}\left(d + \ln\left(\frac{9 \cdot 18}{\delta}\right)\right). \qquad (4.4) \end{aligned}$$

Next, fix an $i \in \{1,2,3\}$ and write $\{(\tilde{X}_{i,1}, \tilde{Y}_{i,1}), \ldots, (\tilde{X}_{i,N_i}, \tilde{Y}_{i,N_i})\} := S_i \cap (ER(h_i) \times \mathscr{Y})$. As h_i and S_i are independent, $\tilde{X}_{i,1}, \ldots, \tilde{X}_{i,N_i}$ are conditionally independent given h_i and N_i. Therefore, we can apply the error bound (4.2) for our base learner L under the conditional distribution given h_i and N_i to conclude: There exists an event E_i' of probability $\geq 1 - \frac{\delta}{9}$ s.t., if $N_i > 0$, then the output h of the base learner L upon input of $S_i \cap (ER(h_i) \times \mathscr{Y})$ satisfies

$$\mathrm{err}_{\mu(\cdot|ER(h_i))}(h, f^*) \leq \frac{8(1+C(\eta_b))}{N_i}\left(d \log \left(\frac{2eN_i}{d(1+C(\eta_b))}\right) + \log\left(\frac{18}{\delta}\right)\right).$$

In particular, on E_i' (if $N_i > 0$) every $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L\left(\mathbb{A}(S_0; T_j)\right)$ satisfies

$$\begin{aligned}
&\mu[ER(h) \cap ER(h_i)] \\
&= \mu[ER(h_i)]\mu[ER(h)|ER(h_i)] && (4.5)\\
&= \mu[ER(h_i)]\mathrm{err}_{\mu(\cdot|ER(h_i))}(h, f^*) && (4.6)\\
&\leq \mu[ER(h_i)]\frac{8(1+C(\eta_b))}{N_i} \\
&\quad \times \left(d \log\left(\frac{2eN_i}{d(1+C(\eta_b))}\right) + \log\left(\frac{18}{\delta}\right)\right). && (4.7)
\end{aligned}$$

Using Chernoff bounds we get that there exists an event E_i'' of probability $\geq 1 - \frac{\delta}{9}$ s.t., if $\mu[ER(h_i)] \geq \frac{2(\frac{10}{3})^2}{\lfloor \frac{m}{4} \rfloor} \ln\left(\frac{9}{\delta}\right)$, then $N_i \geq \frac{7}{10}\mu[ER(h_i)]\lfloor \frac{m}{4} \rfloor$. In particular, on E_i'' we have the implication

$$\mu[ER(h_i)] \geq \frac{2(\frac{10}{3})^2}{\lfloor \frac{m}{4} \rfloor} \ln\left(\frac{9}{\delta}\right) \Rightarrow N_i > 0.$$

If we now combine this with (4.4) and (4.7), then we see: On $E_i \cap E_i' \cap E_i''$, if $\mu[ER(h_i)] \geq \frac{2(\frac{10}{3})^2}{\lfloor \frac{m}{4} \rfloor} \ln\left(\frac{9}{\delta}\right)$, then every $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L\left(\mathbb{A}(S_0; T_j)\right)$ satisfies

$$\begin{aligned}
&\mu[ER(h) \cap ER(h_i)] \\
&\leq \frac{80 \cdot C(\eta_b)}{7\lfloor \frac{m}{4} \rfloor}\left(d \log\left(\frac{2e \cdot \frac{7}{10} \cdot \mu[ER(h_i)]\lfloor \frac{m}{4} \rfloor}{dC(\eta_b)}\right) + \log\left(\frac{18}{\delta}\right)\right) \\
&\leq \frac{80 \cdot C(\eta_b)}{7\lfloor \frac{m}{4} \rfloor}\left(d \log\left(\frac{\frac{7e}{5} \cdot c\left(d + \ln\left(\frac{9 \cdot 18}{\delta}\right)\right)}{d}\right) + \log\left(\frac{18}{\delta}\right)\right) \\
&\leq \frac{80 \cdot C(\eta_b)}{7\lfloor \frac{m}{4} \rfloor}\left(d \log\left(\frac{2}{5}c\left(\frac{7}{2}e + \frac{7e}{d}\ln\left(\frac{18}{\delta}\right)\right)\right) + \log\left(\frac{18}{\delta}\right)\right) \\
&\leq \frac{80 \cdot C(\eta_b)}{7\ln(2)\lfloor \frac{m}{4} \rfloor}\left(d \ln\left(\frac{9ec}{5}\right) + 8\ln\left(\frac{18}{\delta}\right)\right),
\end{aligned}$$

where the last step uses the technical Lemma 5 from the Appendix of Hanneke (2016). As $m > C(\eta_b)c \cdot \ln(18e) - 1 > 3200$, we have $\lfloor \frac{m}{4} \rfloor > \frac{m-4}{4} > \frac{799}{800}\frac{m}{4} > \frac{799}{800}\frac{3200}{3201}\frac{m+1}{4}$. We use this relaxation and compute the logarithmic factors to obtain from the above that

$$\mu[ER(h) \cap ER(h_i)] \leq \frac{600 \cdot C(\eta_b)}{m+1}\left(d + \ln\left(\frac{18}{\delta}\right)\right).$$

Moreover, if $\mu[ER(h_i)] < \frac{23}{\lfloor \frac{m}{4} \rfloor} \ln\left(\frac{9}{\delta}\right)$, then simply because μ is a probability measure, we conclude

$$\begin{aligned}
\mu[ER(h) \cap ER(h_i)] &\leq \mu[ER(h_i)] < \frac{23}{\lfloor \frac{m}{4} \rfloor} \ln\left(\frac{9}{\delta}\right) \\
&< \frac{600 \cdot C(\eta_b)}{m+1}\left(d + \ln\left(\frac{18}{\delta}\right)\right).
\end{aligned}$$

Hence, no matter what value $\mu[ER(h_i)]$ takes, on the event $E_i \cap E_i' \cap E_i''$ we have for all $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L\left(\mathbb{A}(S_0; T_j)\right)$ that

$$\mu[ER(h) \cap ER(h_i)] \leq \frac{600 \cdot C(\eta_b)}{m+1}\left(d + \ln\left(\frac{18}{\delta}\right)\right).$$

Now denote $h_{\mathrm{maj}} = \hat{h}_{m,T} = \mathrm{Majority}(L(\mathbb{A}(S; T)))$ for $S = S_{1:m}$. By definition of the majority function, for any $x \in \mathscr{X}$ at least $\frac{1}{2}$ of the classifiers h in the sequence $L(\mathbb{A}(S; T))$ satisfy $h(x) = h_{\mathrm{maj}}(x)$. So by the strong form of the pigeon hole principle, there exists an $i \in \{1, 2, 3\}$ s.t. $h_i(x) = h_{\mathrm{maj}}(x)$. Also, since each $\mathbb{A}(S_0; T_j)$ contributes an equal number of entries to $\mathbb{A}(S; T)$, for each $i \in \{1, 2, 3\}$, at least $\frac{1}{4}$ of the classifiers $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L\left(\mathbb{A}(S_0; T_j)\right)$ satisfy $h(x) = h_{\mathrm{maj}}(x)$.

In particular, if I is a random variable independent of the training data and distributed uniformly on $\{1, 2, 3\}$ and if $\tilde{h}$ is a random variable conditionally given I and S uniformly distributed on $\bigcup_{j\in\{1,2,3\}\setminus\{I\}} L\left(\mathbb{A}(S_0; T_j)\right)$, then for any fixed $x \in ER(h_{\mathrm{maj}})$, with conditional probability $\geq \frac{1}{12}$, $h_I(x) = \tilde{h}(x) = h_{\mathrm{maj}}(x)$ and thus $x \in ER(h_I) \cap ER(\tilde{h})$.

Hence, for a random variable $X \sim \mu$ independent of the data, of I and of $\tilde{h}$ we can now conclude

$$\begin{aligned}
&\mathbb{E}[\mu[ER(h_i)] \cap ER(\tilde{h}))|S] \\
&= \mathbb{E}[\mathbb{P}[X \in ER(h_I) \cap ER(\tilde{h})|I, \tilde{h}, S]|S] \\
&= \mathbb{E}[\mathbb{1}_{X \in ER(h_I) \cap ER(\tilde{h})}|S] \\
&= \mathbb{E}[\mathbb{P}[X \in ER(h_I) \cap ER(\tilde{h})|S, X]|S] \\
&\geq \mathbb{E}[\mathbb{P}[X \in ER(h_I) \cap ER(\tilde{h})|S, X]\mathbb{1}_{X \in ER(h_{\mathrm{maj}})}|S] \\
&\geq \mathbb{E}[\frac{1}{12}\mathbb{1}_{X \in ER(h_{\mathrm{maj}})}|S] \\
&\geq \frac{1}{12}\mathrm{err}_\mu(h_{\mathrm{maj}}; f^*).
\end{aligned}$$

So on the event $\bigcap_{i\in\{1,2,3\}} E_i \cap E_i' \cap E_i''$ it holds that

$$\begin{aligned}\mathrm{err}_\mu(h_{\mathrm{maj}};\, f^*) &\leq 12\mathbb{E}[\mu[ER(h_i)\cap ER(\tilde{h})]|S] \\ &\leq 12 \max_{i\in\{1,2,3\}} \max_{j\in\{1,2,3\}\setminus\{i\}} \max_{h\in L(\mathbb{A}(S_0;T_j))} \\ &\quad \times \mu[ER(h_i)\cap ER(h)] \\ &< \frac{7200\cdot C(\eta_b)}{m+1}\left(d+\ln\left(\frac{18}{\delta}\right)\right) \\ &= \frac{c\cdot C(\eta_b)}{m+1}\left(d+\ln\left(\frac{18}{\delta}\right)\right).\end{aligned}$$

Since by the union bound the event $\bigcap_{i\in\{1,2,3\}} E_i \cap E_i' \cap E_i''$ has probability $\geq 1-\delta$, the induction step is complete.

It remains to use the claim just proven by induction to derive the desired sample complexity upper bound. For this, take $T = \emptyset$ and note that for $m \geq \lfloor \frac{cC(\eta)}{\varepsilon}\left(d+\ln\left(\frac{18}{\delta}\right)\right)\rfloor$ the right-hand side of (4.3) is $\leq \varepsilon$. Therefore, such a sample size suffices for successful learning using Majority$(L(\mathbb{A}(\cdot;\emptyset)))$. Now recall the discussion before the Theorem, where we observed that $C(\eta_b) \leq \frac{4}{(1-2\eta_b)^2}$, to finish the proof. □

Acknowledgements Open Access funding enabled and organized by Projekt DEAL. M.C.C. wants to thank Michael M. Wolf for suggesting this problem, Gael Sentís and Otfried Gühne for the opportunity to present and discuss the ideas of this paper at the University of Siegen, Srinivasan Arunachalam for his detailed feedback on an earlier draft, and Benedikt Graswald for discussions leading to Example 1. Also, M.C.C. thanks the anonymous reviewers at QTML 2020 and at Springer Quantum Machine Intelligence for their suggestions.

Support from the TopMath Graduate Center of TUM the Graduate School at the Technische Universität München, Germany, from the TopMath Program at the Elite Network of Bavaria, and from the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes) is gratefully acknowledged.

References

Aaronson S (2007) The learnability of quantum states. Proc Roy Soc A Math Phys Eng Sci 463(2088):3089–3114. https://doi.org/10.1098/rspa.2007.0113

Aaronson S (2018) Shadow tomography of quantum states. In: Proceedings of the 50th annual ACM SIGACT symposium on theory of computing, vol 2018. Association for Computing Machinery, New York, pp 325–338. STOC. https://doi.org/10.1145/3188745.3188802

Aaronson S, Hazan EE, Chen X, Kale S, Nayak A (2018) Online learning of quantum states. In: Advances in neural information processing systems, pp 8962–8972

Angluin D, Laird P (1988) Learning from noisy examples. Mach Learn 2(4):343–370. https://doi.org/10.1023/A:1022873112823

Arunachalam S, de Wolf R (2017) Guest column: A survey of quantum learning theory. SIGACT News 48. https://doi.org/10.1145/3106700.3106710. https://pure.uva.nl/ws/files/25255496/p41_arunachalam.pdf

Arunachalam S, de Wolf R (2018) Optimal quantum sample complexity of learning algorithms. J Mach Learn Res 19(71):1–36. http://jmlr.org/papers/v19/18-195.html

Arunachalam S, Chakraborty S, Lee T, Paraashar M, de Wolf R (2019a) Two new results about quantum exact learning. In: Baier C, Chatzigiannakis I, Flocchini P, Leonardi S (eds) 46th International colloquium on automata, languages, and programming (ICALP 2019), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, Leibniz International Proceedings in Informatics (LIPIcs), vol 132, pp 16:1–16:15. https://doi.org/10.4230/LIPIcs.ICALP.2019.16. http://drops.dagstuhl.de/opus/volltexte/2019/10592

Arunachalam S, Grilo AB, Sundaram A (2019b) Quantum hardness of learning shallow classical circuits. arXiv:1903.02840

Arunachalam S, Grilo AB, Yuen H (2020) Quantum statistical query learning

Aslam JA, Decatur SE (1996) On the sample complexity of noise-tolerant learning. Inf Process Lett 57(4):189–195. https://doi.org/10.1016/0020-0190(96)00006-3

Atıcı A, Servedio RA (2007) Quantum algorithms for learning and testing juntas. Quantum Inf Process 6(5):323–348. https://doi.org/10.1007/s11128-007-0061-6

Bartlett PL, Mendelson S (2002) Rademacher and gaussian complexities: Risk bounds and structural results. J Mach Learn Res 3(Nov):463–482. http://www.jmlr.org/papers/volume3/bartlett02a/bartlett02a.pdf

Ben-David S, Dichterman E (1998) Learning with restricted focus of attention. J Comput Syst Sci 56(3):277–298. https://doi.org/10.1006/jcss.1998.1569

Bernstein E, Vazirani U (1993) Quantum complexity theory. In: Kosaraju R (ed) Proceedings of the twenty-fifth annual ACM symposium on Theory of computing. ACM, New York, pp 11–20. https://doi.org/10.1145/167088.167097

Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1989) Learnability and the vapnik-chervonenkis dimension. J ACM (JACM), 36 4 929–965. https://doi.org/10.1145/76359.76371. http://dl.acm.org/ft_gateway.cfm?id=76371&type=pdf

Brandão FGSL, Kastoryano MJ (2019) Finite correlation length implies efficient preparation of quantum thermal states. Commun Math Phys 365(1):1–16. https://doi.org/10.1007/s00220-018-3150-8

Bshouty NH, Jackson JC (1998) Learning dnf over the uniform distribution using a quantum example oracle. SIAM J Comput 28(3):1136–1153. https://doi.org/10.1137/S0097539795293123

Caro MC (2020) Quantum learning boolean linear functions w.r.t. product distributions. Quantum Inf Process 19(6):1–41. https://doi.org/10.1007/s11128-020-02661-1

Cesa-Bianchi N, Dichterman E, Fischer P, Shamir E, Simon HU (1999) Sample-efficient strategies for learning in the presence of noise. J ACM (JACM) 46(5):684–719. https://doi.org/10.1145/324133.324221

Cheng HC, Hsieh MH, Yeh PC (2016) The learnability of unknown quantum measurements. Quantum Inf Comput 16(7-8):615–656
Chowdhury AN (2020) Low, GH, A variational quantum algorithm for preparing quantum gibbs states, Wiebe N
Chung KM, Lin HH (2018) Sample efficient algorithms for learning quantum channels in pac model and the approximate state discrimination problem
Ciliberto C, Herbster M, Ialongo AD, Pontil M, Rocchetto A, Severini S, Wossnig L (2018) Quantum machine learning: A classical perspective. Proc Roy Soc A Math Phys Eng Sci 474(2209):20170551. https://doi.org/10.1098/rspa.2017.0551
Cross AW, Smith G, Smolin JA (2015) Quantum learning robust against noise. Phys Rev A 92(1):97. https://doi.org/10.1103/PhysRevA.92.012327
Grilo AB, Kerenidis I, Zijlstra T (2017) Learning with errors is easy with quantum samples. arXiv:1702.08255
Hanneke S (2016) The optimal sample complexity of pac learning. J Mach Learn Res 17(1):1319–1333. http://dl.acm.org/ft_gateway.cfm?id=2946683&type=pdf
Heinosaari T, Ziman M (2012) The mathematical language of quantum theory: From uncertainty to entanglement. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139031103
Holevo A (1973) Statistical decision theory for quantum systems. J Multivar Anal 3(4):337–394. https://doi.org/10.1016/0047-259X(73)90028-6
Kanade V, Rocchetto A, Severini S (2019) Learning dnfs under product distributions via μ-biased quantum fourier sampling. Quantum Inf Comput 19(15&16):1261–1278. http://www.rintonpress.com/xxqic19/qic-19-1516/1261-1278.pdf
Laird PD (1988) Learning from good and bad data. The Kluwer international series in engineering and computer sciences, knowledge representation, learning and expert systems, vol 47. Springer, Boston. https://doi.org/10.1007/978-1-4613-1685-5
Montanaro A (2012) The quantum query complexity of learning multilinear polynomials. Inf Process Lett 112(11):438–442. https://doi.org/10.1016/j.ipl.2012.03.002
Natarajan N, Dhillon IS, Ravikumar P, Tewari A (2013) Learning with noisy labels, pp 1196–1204
Nielsen MA, Chuang IL (2009) Quantum computation and quantum information, 10th edn. Cambridge Univ. Press, Cambridge
Ristè D, da Silva MP, Ryan CA, Cross AW, Córcoles AD, Smolin JA, Gambetta JM, Chow JM, Johnson BR (2017) Demonstration of quantum advantage in machine learning. Npj Quantum Inf 3(1):16. https://doi.org/10.1038/s41534-017-0017-3
Servedio RA, Gortler SJ (2004) Equivalences and separations between quantum and classical learnability. SIAM J Comput 33(5):1067–1092. https://doi.org/10.1137/S0097539704412910
Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: From theory to algorithms. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781107298019
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142. https://doi.org/10.1145/1968.1972
Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab its Appl 16(2):264–280. https://doi.org/10.1137/1116025
Vershynin R (2018) High-dimensional probability: An introduction with applications in data science, Cambridge series in statistical and probabilistic mathematics vol 47. Cambridge University Press, Cambridge
Watrous J (2018) The theory of quantum information. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781316848142
Wilde M (2013) Quantum information theory. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139525343
Yuen H, Kennedy R, Lax M (1975) Optimum testing of multiple hypotheses in quantum detection theory. IEEE Trans Inf Theory 21(2):125–134. https://doi.org/10.1109/TIT.1975.1055351

A.4 Encoding-dependent generalization bounds for parametrized quantum circuits

Encoding-dependent generalization bounds for parametrized quantum circuits

Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke

---

When using parametrized quantum circuits (PQCs) for variational quantum machine learning (QML) with classical data, the classical input has to be encoded into the PQC. Much of the prior work on variational QML has resolved this issue by first using a quantum feature map to encode the classical input into a quantum state, and then processing this state with trainable quantum gates. However, recent work has demonstrated that distributing data-encoding gates throughout the PQC, instead of just placing them at the initial layer, can have a significant effect on the approximation capabilities of a PQC-based QML model. In this article, we study the influence of such more flexible classical-to-quantum data-encoding strategies on the generalization behaviour of the QML model implemented by the PQC. We prove the fist generalization bounds for PQC-based variational QML that depend explicitly on the data-encoding strategy.

After introducing and motivating our work in Section 1, we recall generalization bounds as a central topic in the classical theory of machine learning in Section 2, with an emphasis on its relevance to model selection. We use Section 3 to introduce the function classes implemented by PQCs that are the main object of study in our work (Eq. (14)). Moreover, we observe that we can represent these functions in terms of generalized trigonometric polynomials (GTPs), in which the achievable frequency spectrum is determined by the classical-to-quantum data-encoding Hamiltonians appearing in the PQC (Eqs. (16) and (30)). We prove this by successively expanding the relevant quantum states in the eigenbases of the encoding Hamiltonians.

Section 4 gives a detailed review of prior work on generalization guarantees for variational QML. With both the relation between PQCs and GTPs and the context provided by prior work established, Section 5 contains our main technical results, namely generalization bounds for GTP function classes. We present two proof strategies towards this goal. First, in Subsection 5.1, we show how to derive upper bounds on the Rademacher complexity of a class of GTPs in terms of the accessible frequency spectrum from known Rademacher complexity bounds for classical feed-forward neural networks (Lemmas 4 and 5). Employing known results from statistical learning theory, these Rademacher complexity bounds imply generalization bounds (Theorem 6). Second, in Subsection 5.2, we use that the size of the accessible frequency spectrum bounds the effective dimensionality of a class of GTPs to prove upper bounds on the covering numbers of such a class (Lemma 9). Through Dudley's entropy integral, these imply Rademacher complexity bounds, which again yield generalization bounds (Theorem 10).

In Section 6, we combine the insights of Sections 3 and 5 to conclude that we can prove generalization bounds for PQC-based QML models by bounding the number of frequency vectors accessible by the encoding strategy. We show how to understand the latter task as a combinatorial question about the spectra of the encoding Hamiltonians. We explore the implications of this reframing of the problem in detail for different practically relevant encoding strategies, which allows us to derive explicitly encoding-dependent generalization bounds for PQC-based models employing these encoding strategies (Corollary 13).

We conclude our paper with Sections 7 and 8, in which we discuss implications of our results. In particular, we emphasize that our results are complementary to much of the prior work on generalization in variational QML, suggest multi-dimensional structural risk minimization as a way of combining these two complementary perspectives for the design of PQCs, and outline questions for future research.

I was significantly involved in finding the ideas and carrying out the scientific work of all parts of this article. The idea for this project arose in discussions between Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, Ryan Sweke, and myself. I developed the proof strategies for our main technical results in Section 5, I was in charge of writing Sections 4 and 5, and I significantly contributed to writing Sections 2 and 7, based on first drafts by Ryan Sweke. Section 1 was written mainly by Elies Gil-Fuster and Jens Eisert. Sections 3 and 6 were mainly written by Johannes Jakob Meyer and Ryan Sweke.

Permission to include:

Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke.
Encoding-dependent generalization bounds for parametrized quantum circuits.
Quantum 5, 582 (2021). https://doi.org/10.22331/q-2021-11-17-582.

Terms and conditions

By submitting a manuscript to Quantum, you agree with Quantum's terms and conditions. In particular, you certify that:

- you have the permission of all co-authors and other right holders to pursue publication of the work in Quantum,
- you are not infringing on anyone's copyright with the material contained in your work,
- you will be fully liable for any charges resulting from copyright infringement, and
- you will not submit this work to any other publishing venue unless it is terminally rejected by Quantum.

In addition, authors, referees and members of all boards of Quantum commit to follow the Code of Conduct laid out here.

The above summary is just for informative purposes. The binding terms and conditions follow.

Table of contents

1. Preamble and definitions
2. Code of conduct
3. Submission and publication of works
4. Data protection and privacy policy
5. Further aspects

1. Preamble and definitions

Access to Quantum is subject to the following terms and conditions, which constitute a contract between Quantum and any user. By engaging in any form of interaction with Quantum or its members, the user accepts these terms and conditions in full.

We denote by "Quantum" the Association for the Promotion of Open Access Publishing in Quantum Science, legally known as

Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften

Müllnergasse 26, 1090 Wien, Austria

(ZVR-Zahl 941922539),

the journal Quantum, the website quantum-journal.org, and the online system provided to organize submission and peer review of works, as well as Quantum's social media accounts.

We denote by "user" any person accessing, using, exchanging, downloading, or submitting any type of content or information with, from and to Quantum and all services it provides, interacting with Quantum in any way, or holding or exercising any function within Quantum.

A "work" submitted to Quantum refers to the actual manuscript as publically available on arxiv.org ("the arXiv"), as well as all supporting material, such as, but not limited to, appendices, supplementary information, a popular summary, datasets, computer code, images, plots, videos or other recordings that are transmitted or made available to Quantum in order to assess the manuscript's suitability for publication. This refers to both the manuscript and material present in the initial submission, as well as all further versions and additions of material in later resubmissions.

The "submitter" is anyone who carries out the submission of a work to Quantum for publication, either through the provided online system or via email, and who is identified by their account in the online system or name and address used in the email. In case of multiple submitters, the definition applies in full to every single one of them.

2. Code of conduct

Quantum fosters scientific integrity and ethical conduct. Consistent with the bylaws of Quantum (such as the constitution of the Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften), its code of conduct upholds those values, detailing the ethical guidelines and expectations for participation in Quantum. Authors, referees, editors, all board members and all other users of Quantum are expected to act at all times in accordance with the principles and standards described in this code.

Unacceptable behaviour

In particular, the code of conduct specifies behaviour that Quantum deems unacceptable, both in interactions with Quantum and in professional life in general. This includes but is not limited to:

1. Plagiarism and fabrication of data and results, including misrepresentation of contributions and authorship, selective reporting, failure to promptly correct errors, or theft of data and/or other research materials, as well as misrepresentation and overstatement of results and the omission of crucial conditions and assumptions.
2. Publication (or submission for publication) of works submitted to or published in Quantum to other publishing venues, unless the work is terminally rejected by Quantum. "Other publishing venues" include other journals and conference proceedings, but exclude public pre-print servers such as the arXiv and personal or institutional websites of the authors. Re-publication of excerpts or the entirety of a work submitted to or published in Quantum as part of a work with a broader scope, such as a review article of thesis, is explicit allowed.
3. Subversion of peer review, including failure to declare conflict of interest, failure

to recuse under conflict of interest, misuse of information during review, unnecessarily delaying the peer-review process, violation of the anonymity of referees, premature solicitation of press coverage, corruption and/or bribery.

4. Impersonation of other persons or entities, as well as unrightfully claiming the ownership of scientific titles, professional positions, or affiliations.
5. Using Quantum or any other system for the dissemination of scientific works to promote hate or discriminatory speech, or to infringe on the rights of others.
6. Sharing of confidential information, such as the identity of reviewers, referee reports, and other internal correspondence to persons not involved in the peer-review process.
7. Discrimination of any kind, such as on the basis of religion, disability, age, national origin, race, ethnicity, sexual orientation, gender identity, or gender expression. Discrimination includes the use of derogatory comments or slurs.
8. Harassment, including bullying and intimidation, false accusations, threats and assault, as well as sexual harassment in public or in private.
9. The violation of public trust, including making false or misleading statements, to media, and misrepresentation to grant and/or funding agencies.

Reporting and investigation

Quantum may learn of violations through reports from witnesses or aggrieved individuals and parties, including anonymous reports filed through the dedicated online form. For the safety of all reporters, once a report has been made, Quantum editors and board members are bound to maintain the confidentiality of the report except as explicitly requested by the reporting parties. Upon receiving a report, Quantum will progress as follows:

1. The Executive Board of Quantum will name an investigator or form a small investigation body of no more than three people, each of whom must be free of conflict of interest.
2. The investigator or investigating body may solicit additional information from the reporter, with the goal of reaching a tentative conclusion over the course of two weeks.
3. The tentative conclusion of the investigating body will be delivered to the Steering Board, along with a suggested resolution action as described in the section below.
4. If the tentative conclusion and suggested resolution action are agreed upon by the Steering Board, Quantum will inform the reporter of their decision and seek agreement before proceeding.
5. The party suspected of a violation of the code of conduct will be informed of the allegations and planned resolution action and given 20 working days to respond.
6. Depending on the findings, on communication from the involved parties, and consensus of the Steering Board, the resolution action may be implemented or further investigations carried out with the aim of resolving the situation.

During the reporting and investigating process, all individuals must exercise all due diligence to prevent divulging any report details beyond those strictly necessary to enact and uphold the code of conduct. In particular, if upon receiving an initial report, it is deemed the alleged infraction would not result in a penalty more severe than a formal warning, the Executive Board may decide to directly handle the report without the aid of an investigating body, provided that no conflict of interest is introduced.

Enforcement and penalties

If a Quantum user is found through the preceding process to have committed any violation of the code of conduct, Quantum may enforce the code of conduct in a number of different ways. An appropriate resolution is decided by the Steering Board, taking into account all factors, and having as a goal to improve the situation. Possible actions to enforce the code of conduct include but are not limited to:

1. A formal (written) warning made to the infringing party.
2. Requiring the infringing party to make a formal (written) apology.
3. Reporting the infringing party to their home institutions, employers, and/or professional societies.
4. Reporting the infringing party to the relevant authorities, in case of suspicion of criminal offences.
5. Retraction of compromised manuscripts (based on scientific reasons).
6. Refusal to consider future manuscripts from the infringing party.
7. Expulsion from the Steering, Executive or Editorial Board.

3. Submission and publication of works

Works submitted to Quantum undergo the peer-review process following the Editorial Policies of Quantum. This process ends with either the acceptance of the work for publication, or the terminal rejection of the work.

Responsibilities of the submitter

By submitting a work to Quantum, the submitter warrants all of the following points and assumes full responsibility and liability for any costs and damages resulting directly or indirectly from any of them being untrue:

1. The submitted work is an original creation of the authors listed on the manuscript, and all listed authors have made substantial contributions to the creation of the work.
2. The submitter has the permission of all authors and all other copyright and intellectual property rights holders to pursue the publication of the work in Quantum, and to grant Quantum all the rights specified in these terms and conditions.
3. The manuscript is publicly accessible on the arXiv in the section quant-ph, or at

least crosslisted to quant-ph.

4. The work has not been previously published in any other journal or publishing venue, except in conference proceedings and on public pre-print servers such as the arXiv or the authors' personal or institutional websites. Works previously published in conference proceedings must substantially differ from or expand upon the conference version (for example contain previously omitted proofs) and indicate the previous publication on the first page of the manuscript.
5. The submitter has obtained permissions to grant Quantum the rights specified in these terms and conditions for all material contained in the work and has included appropriate credits and prominently marked or indicated any rights held by third parties.
6. The submitter has clearly informed Quantum at the time of submission of any parts of the work which, due to copyright or other constraints, cannot be published by Quantum under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.
7. In case of acceptance, the final published version of the work will be uploaded on the arXiv under the Creative Commons Attribution 4.0 International (CC BY 4.0).
8. In case of acceptance, the final published version of the work on the arXiv complies with the Crossref DOI guidelines. In particular, all references cited by the submitted work that have a DOI assigned to them contain DOI links.
9. In case of acceptance and publication in Quantum, the work will not be submitted to other publishing venues, such as journals or conference proceedings.

Rights of the submitter

The submitter is granted the following rights:

1. At any point prior to acceptance the submitter, as well as any author of the work, can withdraw a work from Quantum. A notification of withdrawal has to be submitted to the handling editor either through the online submission system or by email. Upon receiving a notification of withdrawal prior to acceptance, Quantum terminally rejects the work, thereby ending the peer-review process.
2. The submitter, as well as any author of the work, can also withdraw a work from Quantum after acceptance and publication by notifying Quantum through email. This however does not trigger a terminal rejection of the work and in particular does not invalidate the rights granted to Quantum during submission. Quantum will instead put a notification on the publication page that the work was withdrawn.

Internal correspondence

Unless explicitly agreed otherwise by all parts involved, all correspondence between editors, referees and authors during the peer review process should be treated as confidential, and may only be shared with the present and future Editorial Board

members who have not declared a conflict of interest with the work.

Rights of Quantum

By submitting a work to Quantum, the submitter explicitly grants Quantum the following additional rights:

1. The right to terminally reject the work, in particular on the basis of the editors' judgement and/or referee reports.
2. The right to permanently store and share the work, referee reports, and intermediate correspondence with the referees and all current and future members of the Editorial Board who have not declared a conflict of Interest.
3. The right to share the identity of the referees with all members of the current and future Editorial Boards who have not declared a conflict of Interest.
4. The non-exclusive right to share, publish, host, distribute, print, advertise, classify, and otherwise use the manuscript, other parts of the work and all metadata associated with it under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence, unless the work is terminally rejected by Quantum, expect for parts of the work that are covered by incompatible licences. This does not imply any restrictions on the right to publish parts or the entirety of the work of other parties.
5. The right to deposit the metadata associated with the work in the Crossref system and to assign a DOI to the work.
6. The right to publish anonymized statistics on submissions and the peer-review process.

Copyright of works published by Quantum

All manuscripts and other parts of works that were previously submitted to Quantum and then published by Quantum, as well as the associated meta-data, including for example a work's title, abstract, author list, figures, datasets, or popular summary, are published under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

For material associated with a manuscript, such as that linked to from the manuscript, or a work's page on Quantum's website, especially if hosted on other platforms, other licences can apply.

Each owner of copyright on parts or the entirety of a work submitted to or published by Quantum retains their copyright.

4. Data protection and privacy policy

The purpose of this data protection policy is to inform all users of Quantum and this website about the type of personal data that is collected and processed by Quantum and the company providing the hosting infrastructure for this website, as well as the purpose of said data processing.
Quantum is taking data protection very seriously, and it is treating your personal data according to the legal requirements.
In particular, Quantum complies with the European General Data Protection Regulation

(GDPR).
Please keep in mind that any data transmission over the Internet and any form of digital data processing can be affected by security flaws is software and hardware that are beyond the control of Quantum.
A complete protection from unauthorized access by third parties can thus never be fully guaranteed.

As Quantum does not fall in any of the categories of entities that are required to appoint a DPO according to the GDPR, it does not have a dedicated DPO.
Should you have any questions on this data protection and privacy policy, please contact us through one of the channels described in the impressum.

Personal data

Personal data is any information related to a natural person or data subject that can be used to directly or indirectly identify the person.
This website collects personal data only to the extent necessary and in a way that is legally permissible (see below for more details).

Cookies

This website uses cookies. A cookie is a small file that is saved on the device with which you are accessing this website. Should you not want to be served a cookie when using this website, most common browsers can be configured to disallow the usage of cookies. This may affect the usability of this site (such as the ability to opt out of the data collection for analytics, see below).

Sharing buttons

The sharing buttons for sharing content on social media and other platforms displayed on this website are provided by the WordPress plugin Shariff Wrapper. By design, the Shariff Wrapper plug-in does not transmit any data to the social media platforms and other sharing services unless one of the sharing buttons is explicitly clicked. Shariff Wrapper's statistics feature is disabled on this site, so that no personal data about sharing activity is stored.

Data stored and processed of all users

Quantum and the company hosting this website may collect, store, and process the following data of all users of Quantum:

1. visited pages
2. time of access
3. number of transmitted bytes
4. link that lead to the page being accessed
5. browser used
6. operating system used
7. ip address from which it was accessed

8. data (text, files, tick boxes, ...) entered in forms and search boxes

The data may be collected in server log files and for the purpose of analyzing how this website is being used (analytics) by means of the WordPress plug-in WP statistics.
Unless stated otherwise below, the data is saved exclusively on the servers of the company providing the hosting infrastructure for this website and on computers controlled by Quantum (for the purpose of archiving and backup) and can only be accessed through password protected accounts on these systems and is not shared with any third party service such as Google Analytics.

Unless stated otherwise below, the data collected in this way is used exclusively for statistical analysis and improvement of the website as well as to ensure its safe and lawful operation.
In particular, ip addresses are saved only in pseudo-anonymized form, either hashed or with the last block truncated.

If the WP statistics plug-in is enabled, you have the option to opt-out of the data collection for analytics by performing the following steps: (1) Clear all cookies for this website in your browser. (2) Refresh the page. (3) Click/tap the button “opt out” in the banner at the bottom of the page.

Data entered into or resulting from the use of forms and transmitted via email

Data entered into or resulting from the use of forms on this website and such sent to Quantum via email may be processed in additional ways.
In particular, it may be stored and processed on information technological devices controlled by members of Quantum as well as in password protected collaborative working and cloud computing platforms such as Jira, Dropbox, Google Drive, and Google Docs/Google Sheets.
For data relevant for the peer-review and publication process additional rules apply (see the next section for more details).
If the data entered in such forms is stored or processed of purposes other than those described in the last section, the form contains a more detailed description of the type and purpose of the data storage and processing.
You have to tick a tick box on such forms to express your explicit consent to such additional data storage and processing.

Data stored and processed for and during the peer-review and publication process

During the peer-review and publication process (which includes the handling of appeals) additional personal data is collected, stored, and processed by Quantum, as well as the third-party service < href="https://scholasticahq.com/">Scholastica, namely:

1. written communication with and between authors, editors, and referees
2. the times and other meta-data associated with this communication
3. the email addresses and account names at other services used to carry out this communication

4. all data and material provided to Quantum by the submitter, including the submitted work

This this data is permanently stored by Quantum to ensure long-term accountability and justifiability of editorial decisions and to ensure a means of contacting the submitter or authors (e.g., in case of later corrections to published works).
Such data may further be stored and processed on the collaborative working platform Jira and the third-party service < href="https://scholasticahq.com/">Scholastica to the extent this is necessary to carry out and supervise the peer-review process.
On such platforms the data is accessible only through password protected accounts.

Quantum retains the right to process this data for the purpose of performing statistical analysis of the peer-review process, correlate this data with other publicly available information, and to publish such analysis and the underlying data in a suitably anonymized form.
In doing this, Quantum takes the utmost care to ensure that no personally identifying information is leaked and in particular that the anonymity of referees is guaranteed.

For accepted works, Quantum publishes the work itself (including all companioning material and meta-data) in accordance with the terms and conditions.
In particular, as part of its service, Quantum uploads the meta-data of accepted works, as well as parts or all of the work, to platforms such as Crossref, the Directory of Open Access Journals, Clarivate Analytics, and Clockss for the purpose of registering DOIs, making the work discoverable by readers, facilitate biometrics, and archiving.

Data on payments and donations

Data on payments of article processing charges and donations are obviously available to the handling services, such as the involved banks and, if applicable, PayPal. Such data is processed for accounting purposes in accordance with applicable law.
On top of that, Quantum practices public accounting. All donations and payed fees are made available in this publicly shared shared spreadsheet.

Data publicly available on this website

Data that is anyway publicly available on this website, my be share by Quantum in posts on social media platforms including but not limited to Facebook, Twitter, and LinkedIn, either directly of by means of third-party services such as buffer.com or IFTTT in accordance with the license under which this data was published.

Backups

All data described above may be included in backups to ensure the continued availability of the services of Quantum. Such backups are stored “off-site”, i.e., in a different physical location than the production system on servers or storage systems controlled by Quantum or people working for Quantum and access to which is suitably restricted with passwords. The data in backups is not processed and is stored in such a way that it can not even be directly accessed from the production systems.

Data protection

Quantum uses strong and individual passwords for all accounts and services that can be

used to access personal data of its users. Quantum uses individual email accounts for each third party service to ensure a fine grained access to personal data, restricted to only those members of Quantum who actually need access.

Data breach procedures

In case of a data breach that reveals otherwise not publicly available personal data in our own data handling systems or any of the third party services Quantum relies on, all identifiable users will be notified in due time after a suitable assessment of the situation.

User rights

As a user, you have the right to request information about which of your personal data is stored and processed by Quantum and for what purpose. You can demand that personal data is corrected (in case it is incorrect) or deleted (to the extend this is compatible with the terms and conditions of Quantum and applicable law), and have the right to obtain a digital copy of the stored personal data. To contact Quantum on such matters, please use one of the channels described in the impressum.

Data protection and privacy policy changes

Quantum may change its data protection and privacy policy from time to time, at Quantum's sole discretion. Quantum encourages its users to check this page for changes frequently. Your continued use of any service Quantum provides, including but not limited to this website, after any change in this data protection and privacy policy constitutes your acceptance of such change.

5. Further aspects

Disclaimer of warranty

There is no warranty for the services provided by Quantum, to the extent permitted by applicable law. Except when otherwise stated in writing, the copyright holders and/or other parties provide these services "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose or future availability. The entire risk as to the quality and performance of the services is with the user. Should the services prove defective, the user assumes the cost of all damages incurred.

Limitation of liability

In no event, unless required by applicable law or agreed to in writing, will Quantum be liable to any user for damages, including any general, special, incidental or consequential damages arising out of the use or inability to use the services provided by Quantum, even if Quantum has been advised of the possibility of such damages.

External content

Links to other works, websites, or other documents, including but not limited to hyperlinks on the website quantum-journal.org as well as in manuscripts published by Quantum, may link to content that is beyond the control of Quantum. Quantum hence does not assume any kind of responsibility or liability for the content to which such links

point or transmissions that can be received through them.

Infringement notification

In case you believe that any material published by Quantum infringes on your copyright or intellectual property rights in any way, you must contact in writing, in either English or German, the

Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften
Müllnergasse 26, 1090 Wien, Austria.

Logos, images and materials created by Quantum

The term "Quantum" and the Quantum logo are a registered trademark of the Verein zur Förderung des Open Access Publizierens in den Quantenwissenschaften in the European Union (EUIPO) for the publishing of scientific papers, electronic publishing, the publishing of journals, and several other categories of goods and services.

All other company and product names, logos, trademarks, registered trademarks, and brands are property of their respective owners. All such company, product and service names used in this website are for identification purposes only. Use of these names, logos, and brands does not imply endorsement.

The logo of Quantum, as well as all images created by Quantum, are published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) licence. The LaTeX template quantumarticle is published on GitHub under the LaTeX Project Public licence version 1.3c.

Right to modify terms and conditions

Quantum retains the right to modify these terms and conditions following consultation with the Steering Board. All submitters of works undergoing peer-review at the time of change must be notified if affected by the changes. Submitters are always bound to the version of these terms and conditions at the time of the last (re-)submission.

These terms and conditions, as well as the Editorial Policies and the Constitution of Quantum, are published under a Creative Commons "No Rights reserved" (CC0) licence. Sharing, tweaking and reuse of these policies is allowed and encouraged. If you adapt them to other projects, we would love to hear from you.

Encoding-dependent generalization bounds for parametrized quantum circuits

Matthias C. Caro[1,2], Elies Gil-Fuster[3,4], Johannes Jakob Meyer[3], Jens Eisert[3,4,5], and Ryan Sweke[3]

[1]Department of Mathematics, Technical University of Munich, 85748 Garching, Germany
[2]Munich Center for Quantum Science and Technology (MCQST), 80799 Munich, Germany
[3]Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany
[4]Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[5]Helmholtz-Zentrum Berlin für Materialien und Energie, 14109 Berlin, Germany

2021-10-28

A large body of recent work has begun to explore the potential of parametrized quantum circuits (PQCs) as machine learning models, within the framework of hybrid quantum-classical optimization. In particular, theoretical guarantees on the out-of-sample performance of such models, in terms of generalization bounds, have emerged. However, none of these generalization bounds depend explicitly on how the classical input data is encoded into the PQC. We derive generalization bounds for PQC-based models that depend explicitly on the strategy used for data-encoding. These imply bounds on the performance of trained PQC-based models on unseen data. Moreover, our results facilitate the selection of optimal data-encoding strategies via structural risk minimization, a mathematically rigorous framework for model selection. We obtain our generalization bounds by bounding the complexity of PQC-based models as measured by the Rademacher complexity and the metric entropy, two complexity measures from statistical learning theory. To achieve this, we rely on a representation of PQC-based models via trigonometric functions. Our generalization bounds emphasize the importance of well-considered data-encoding strategies for PQC-based models.

1 Introduction

Recent years have witnessed a surge of interest in the question of whether and how quantum computers can meaningfully address computational problems in machine learning [1, 2]. This development has been largely driven by two factors. On the one hand, there is evidence that some quantum machine learning algorithms may lead to an increased performance over classical algorithms for the analysis of classical data with respect to important figures of merit [3–7]. On the other hand, the increasing availability of quantum computational devices provides significant stimulus. While these "noisy intermediate-scale quantum" (NISQ) devices are still a far cry from full-scale fault-tolerant quantum computers, there exists growing evidence that they may be able to out-perform classical computers on some highly-tailored tasks [8]. Given the inherent limitations of NISQ devices, most current approaches to near-term quantum-enhanced machine learning fall under the umbrella of hybrid quantum-classical algorithms [9]. Of particular prominence are variational quantum algorithms in which a *parametrized quantum circuit* (PQC) is used to define a machine learning model which is then updated via a classical optimizer [10–12].

There is a wealth of architectural choices for PQC-based machine learning models. These include the width and depth of the quantum circuit, the precise layout and structure of trainable gates, as well as the mechanism via which classical data is encoded into the quantum circuit. The flexibility in design choices for PQCs is often only perceived strongly in terms of the structure and layout of the trainable gates [13, 14]. However, when using a PQC to define a machine learning model for *classical data*, the data-encoding strategy becomes a necessary architectural

design choice, which has received comparably little attention. Despite this, it has recently been shown that the data-encoding strategy is directly related to the expressive power of PQC-based models [15–17]. In this work, we further the study of data-encoding strategies for PQC-based supervised learning models by investigating the effect of data-encoding strategies on *generalization* performance.

More specifically, we consider the following fundamental question: Given a PQC-based model which has been trained on a specific data set, can we place any guarantees on its expected *out-of-sample* performance, i.e., its expected accuracy on new data, drawn from the same distribution as the training set? This question is motivated by the key insight that one should *not* choose the model or architecture which performs best on the available training data, but rather the model for which one expects the best out-of-sample performance. Typically, one refers to the difference between the accuracy of a model on a given training set and its expected out-of-sample accuracy as the *generalization gap*. We call a (probabilistic) upper bound on this generalization gap a *generalization bound*. Historically, techniques for both proving generalization bounds and for using generalization bounds for principled model selection have been developed under the umbrella of *statistical learning theory* [18–20].

We start by presenting a selection of central notions in statistical learning theory. Of particular interest is the relation between generalization bounds and *complexity measures* of different types. Indeed, due to a large body of existing literature, bounding the generalization gap of a learning model typically reduces to bounding some quantifiable property of the hypothesis class used for learning. There are many examples of such complexity measures (also known as *capacity metrics* or just *expressivity measures*), and based on their specifics they are used for different learning models, either quantum or not. In this work, we employ generalization bounds based on the Rademacher complexity and the metric entropy. However, we want to mention that there are also other important approaches to generalization not taken here, such as stability [21], compression [22], or the PAC-Bayesian framework [23].

Given the fundamental role of generalization bounds, there has recently been a strong and steady stream of works contributing to the derivation of generalization bounds for PQC-based models [24–32]. However, as discussed in detail in Section 4, these prior works all differ from our results in a variety of ways. Firstly, they considered only "encoding-first" PQC architectures, in which the PQC-based models are assumed to consist of an initial data-encoding block, mapping a classical input to a data-dependent quantum state, followed by a circuit consisting only of fixed and trainable gates. In contrast, we consider PQC-based models incorporating *data re-uploading* [17], in which trainable circuit blocks are interleaved with data-encoding circuit blocks. This is particularly relevant given the results of Refs. [15, 33], which have illuminated the significant effects of data re-uploading on the expressive power of PQC-based models.

Additionally, our work is the first to provide a generalization bound from which it is immediately clear how altering the data-encoding strategy influences the generalization performance of the model. This is possible because our bound depends *explicitly* on architectural hyper-parameters associated with the data-encoding strategy. This sets our results apart from prior art where the data-encoding figured only *implicitly*, if at all. We discuss this difference between implicitly and explicitly encoding-dependent generalization bounds more concretely in Section 4.

In order to obtain our generalization bounds, we rely strongly on a representation of PQC-based models via generalized trigonometric polynomials (GTPs), which has been previously derived in Refs. [15, 33]. In particular, we exploit the fact that the data-encoding strategy of the PQC-based model directly determines the frequency spectrum of the corresponding GTPs. As such, the number of accessible frequencies in the GTP representation provides a natural measure of the complexity of a particular data-encoding strategy. Given this, we first derive generalization bounds for GTPs, which exhibit a dependence on the square root of the number of accessible frequencies. We then proceed to determine, for different data-encoding strategies, upper bounds on the number of accessible frequencies in the GTP representation. We use these results to identify a variety of natural data-encoding strategies for which the number of accessible frequencies, and therefore the associated generalization bounds, scale polynomially with the number of data-encoding gates. While one *cannot* use generalization bounds alone to recommend an optimal data-encoding strategy, we discuss how these generalization bounds can be combined with empirical risk estimates, via *structural risk minimization*, to facilitate the selection of an optimal data-encoding strategy

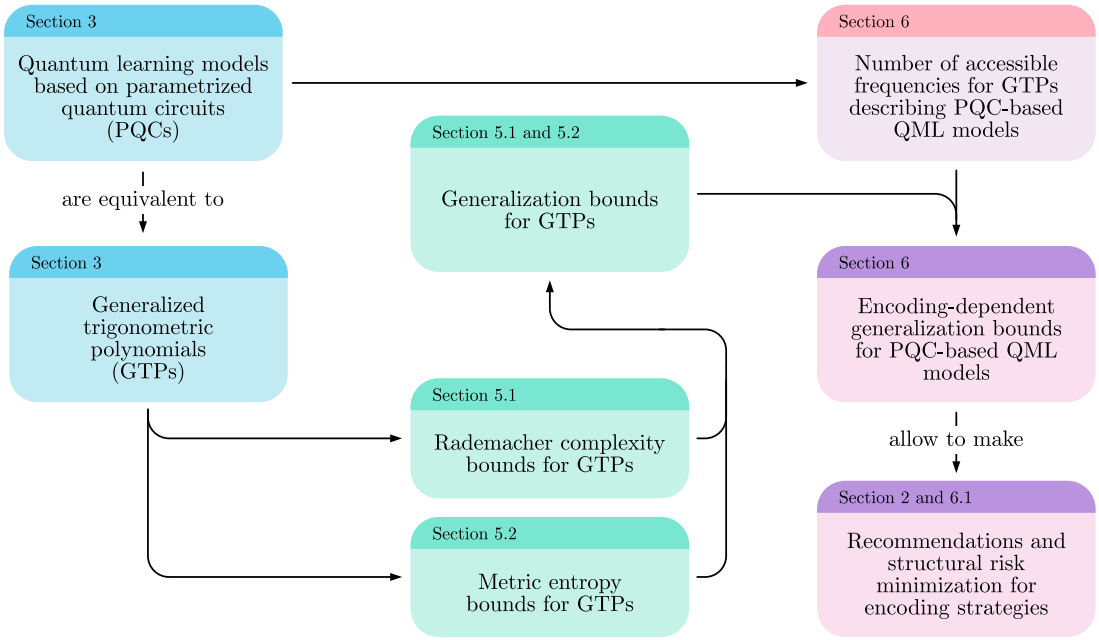


Figure 1: A flowchart of the argument presented in this work.

for a given problem.

1.1 Structure of this work

This work is structured as follows: Section 2 gives a pedagogical introduction to statistical learning theory, explains the importance of generalization bounds, and discusses the structural risk minimization principle. After establishing these concepts, we formulate the main questions addressed in this work. In Section 3, we begin by introducing the PQC-based learning models used in this work. We then present a detailed discussion of the approach of Ref. [33], which demonstrates how the functions implemented by a PQC-based model can be represented by generalized trigonometric polynomials. In particular, we emphasize how the data encoding strategy of the PQC-based model translates to the accessible frequencies of the generalized trigonometric polynomials. Section 4 then provides a detailed review of prior work on generalization in quantum machine learning. In Section 5, we establish generalization bounds for classes of generalized trigonometric polynomials in terms of the number of accessible frequencies. We present one approach via the Rademacher complexity (Section 5.1) and another via covering numbers (Section 5.2). Section 6 then expands upon Section 3 by deriving upper bounds on the number of accessible frequencies, in the generalized trigonometric polynomial representation of the PQC-based models associated with different data-encoding strategies. This analysis allows us to use the results from Section 5 to state explicitly encoding-dependent generalization bounds for PQC-based models, and to compare different encoding strategies from a generalization perspective. We discuss the implications of our results in Section 7. In particular, we emphasize how our results are complementary to many prior works, but also describe how the different approaches can be combined. Additionally, we sketch some directions for future research. Section 8 contains a short summary of our work. The logical flow of this manuscript is visualized in Figure 1.

2 Motivation: Generalization bounds, sample complexities and model selection

To motivate the content of this work and to define the setting, we start with a brief and select introduction to the framework of *statistical learning theory*. Interested readers are referred to Refs. [20] and [34] for a more detailed presentation. Within this framework, any supervised learning

problem is defined by a *domain* $\mathcal{X}$, a *co-domain* $\mathcal{Y}$, a *probability distribution* P over $\mathcal{X} \times \mathcal{Y}$ and a *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. We assume that $\mathcal{X}, \mathcal{Y}$ and ℓ are known, while P is unknown. We will denote the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$ as $\mathcal{Y}^{\mathcal{X}}$. To gain intuition, it is useful to think of the situation in which there exists a deterministic rule for assigning predictions to domain elements. We can model this in the framework outlined above with an unknown target function $f \in \mathcal{Y}^{\mathcal{X}}$, as well as some unknown probability distribution $P_{\mathcal{X}}$ over $\mathcal{X}$, such that samples from P are obtained by first drawing a domain element $\boldsymbol{x} \in \mathcal{X}$ from $P_{\mathcal{X}}$, and then outputting the tuple $(\boldsymbol{x}, f(\boldsymbol{x}))$, i.e.

$$P(\boldsymbol{x}, y) = \begin{cases} P_{\mathcal{X}}(\boldsymbol{x}) & \text{if } y = f(\boldsymbol{x}), \\ 0 & \text{if } y \neq f(\boldsymbol{x}). \end{cases} \tag{1}$$

In general, however, it may be the case that there exists $y_1 \neq y_2$ for which both $P(\boldsymbol{x}, y_1) > 0$ and $P(\boldsymbol{x}, y_2) > 0$, i.e., that the underlying process for labeling data points is not deterministic.

Additionally, we are given a training data set

$$S = \{(\boldsymbol{x}_i, y_i) \sim P \,|\, i \in \{1, \ldots, m\}\} \tag{2}$$

of m tuples drawn independently from (the unknown distribution) P, and our goal is to design a *learning algorithm* $\mathcal{A}$ which, given S as input, outputs a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ that achieves a sufficiently small *risk*

$$R(h) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, h(\boldsymbol{x}))\, \mathrm{d}P(\boldsymbol{x}, y). \tag{3}$$

Informally, we often refer to the risk $R(h)$ as characterizing the *out-of-sample* performance of the hypothesis h, as it is this quantity which tells us how well we can expect the hypothesis h to perform on (possibly previously unseen) future data drawn from P. It is critical to note, however, that as the underlying probability distribution P is unknown, given a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$, one *cannot* directly evaluate $R(h)$. In light of this, a natural alternative is to evaluate the *empirical risk* of h with respect to S, which is defined as the average loss over the training samples

$$\hat{R}_S(h) = \frac{1}{|S|} \sum_{(\boldsymbol{x}_i, y_i) \in S} \ell(y_i, h(\boldsymbol{x}_i)). \tag{4}$$

In contrast to the risk $R(h)$, the empirical risk $\hat{R}(h)$ characterizes the *in-sample* performance of h with respect to the data set S, which has been sampled from P.

Naively, one might hope to be able to construct learning algorithms which could in principle output *any* $h \in \mathcal{Y}^{\mathcal{X}}$. However, the "no-free-lunch" theorem rules out the possibility of meaningful learning in this case [35], and therefore we typically consider learning algorithms whose range is some subset $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$. We then refer to $\mathcal{F}$ as the *hypothesis class* associated with the learning algorithm which is, by assumption, also known to the learning algorithm. To gain some intuition, one could think of $\mathcal{F}$ as the set of all functions realizable by neural networks of some fixed width and depth, or, as we describe in Section 3, as the set of all functions realizable by a parametrized quantum circuit model with some fixed architecture. With respect to this setting, the following natural question arises: Suppose we have a learning algorithm $\mathcal{A}$ with hypothesis class $\mathcal{F}$, which has been run on a randomly drawn data set of m samples $S \sim P^m$ and outputs some hypothesis $h \in \mathcal{F}$, as well as some "training log" which we denote by $\mathrm{hist}(\mathcal{A}, S)$[1]. Given the achieved empirical risk $\hat{R}_S(h)$, can we put an upper bound on the true risk $R(h)$, which holds with high probability over the randomly drawn data set S? More specifically, can we make a statement of the form: For all $\delta \in (0, 1)$, with probability $1 - \delta$ over $S \sim P^m$, for all $h \in \mathcal{F}$ we have that

$$R(h) \leq \hat{R}_S(h) + g(\mathcal{F}, h, m, S, \mathcal{A}, \mathrm{hist}(\mathcal{A}, S), \delta). \tag{5}$$

We refer to such a statement as a *generalization bound*, and note that the function g appearing in Eq. (5) provides a (probabilistic) upper bound on the quantity $R(h) - \hat{R}_S(h)$, which we call *generalization gap* (of h with respect to S). Such bounds are desirable because they allow us to

[1] Such a training log could for example record the value of the empirical risk, or properties of the trial hypotheses (such as weight matrices for neural networks), at each stage of an iterative optimization procedure.

leverage the information we have access to – i.e., the empirical risk, and properties of the learning algorithm, data set and optimization procedure – to upper bound $R(h)$, which is the quantity we do not have access to, but are ultimately interested in. In general, as indicated explicitly in Eq. (5), the upper bound g on the generalization gap could depend on properties of the achieved hypothesis h, properties of the data set S, properties of the learning algorithm $\mathcal{A}$, and details of the optimization that led to h. However, in this work we will focus on *uniform* generalization bounds of the form: for all $\delta \in (0,1)$, with probability $1-\delta$ over $S \sim P^m$, we have for all $h \in \mathcal{F}$ that

$$R(h) \leq \hat{R}_S(h) + g(\mathcal{F}, m, \delta). \tag{6}$$

To be specific, we focus on generalization bounds for which the upper bound on the generalization gap – i.e., the function g – depends only on properties of the hypothesis class $\mathcal{F}$, the data set size m and the desired probability δ. We note that the term "uniform" is used when describing such generalization bounds to indicate that, with respect to a fixed data set size m and probability threshold δ, the upper bound on the generalization gap will be the same – i.e., uniform – for all $h \in \mathcal{F}$. While it is known that there exist scenarios in which uniform generalization bounds are not tight [36, 37], we postpone a discussion of these issues to Section 7.

As motivated above, given a uniform generalization bound for a hypothesis class $\mathcal{F}$, one typical application is as follows: Given a data set S sampled from P, with $|S| = m$, run some learning algorithm to obtain a hypothesis $h \in \mathcal{F}$, evaluate its empirical risk $\hat{R}_S(h)$, and then use the generalization bound to place a (probabilistic) upper bound on the true risk $R(h)$. However, we can also often straightforwardly use such a generalization bound to answer the following natural question: Given some $\epsilon > 0$ and some $\delta \in (0,1)$, what is the minimum size of S sufficient to ensure that, with probability $1-\delta$, for all $h \in \mathcal{F}$, the generalization gap satisfies $R(h) - \hat{R}_S(h) \leq \epsilon$? To see this, note that if we have a uniform generalization bound, then by setting

$$g(\mathcal{F}, m, \delta) \leq \epsilon \tag{7}$$

and solving for m, it is often possible to find some function $f(\epsilon, \delta, \mathcal{F})$ such that, with probability $1-\delta$ over $S \sim P^m$,

$$m \geq f(\epsilon, \delta, \mathcal{F}) \Rightarrow \forall h \in \mathcal{F} : R(h) - \hat{R}_S(h) \leq g(\mathcal{F}, m, \delta) \leq \epsilon. \tag{8}$$

As the generalization bound may not be tight, we therefore see that $f(\epsilon, \delta, \mathcal{F})$ provides an *upper bound* on the minimum size of S sufficient to probabilistically guarantee a generalization gap less than ϵ for all $h \in \mathcal{F}$.

Finally, apart from the fundamental applications of allowing us to bound the out-of-sample performance of a hypothesis, or upper bound the minimum sample-size sufficient to guarantee a certain generalization gap, generalization bounds also allow us to address the issue of *model selection*, via the framework of *structural risk minimization* [20]. Importantly, we note that one *cannot* simply use only the function $g(k, m, \delta)$ for model selection: A trivial learning model, which outputs the same hypothesis independently of the input data, has $g(k, m, \delta) = 0$, but cannot achieve good prediction performance on interesting tasks. Structural risk minimization thus suggests combining a generalization bound with an empirical risk evaluation on a specific data-set to choose the model with the smallest upper-bound on the true risk. More specifically, let us assume that our hypothesis class depends on some "architectural hyper-parameter" k, with some notion of ordering such that

$$k_1 \leq k_2 \implies \mathcal{F}_{k_1} \subseteq \mathcal{F}_{k_2}. \tag{9}$$

For example, $\mathcal{F}_k$ could be the set of all neural networks of fixed width and depth k. Given this, how should we choose the hypothesis class – or model complexity – that we use for a given learning problem? As illustrated in Figure 2, generalization bounds, when combined with empirical risk evaluations, can allow us to answer this question. In particular, assume that we have a uniform generalization bound of the form: For all $\delta \in (0,1)$, with probability $1-\delta$ over $S \sim P^m$, for all $h \in \mathcal{F}_k$,

$$R(h) \leq \hat{R}_S(h) + g(k, m, \delta), \tag{10}$$

where $g(k, m, \delta)$ is non-decreasing with respect to k. Here, we have written $g(k, m, \delta)$ rather than $g(\mathcal{F}_k, m, \delta)$ to emphasize the assumption that the hyper-parameter k is the only property of $\mathcal{F}_k$ on

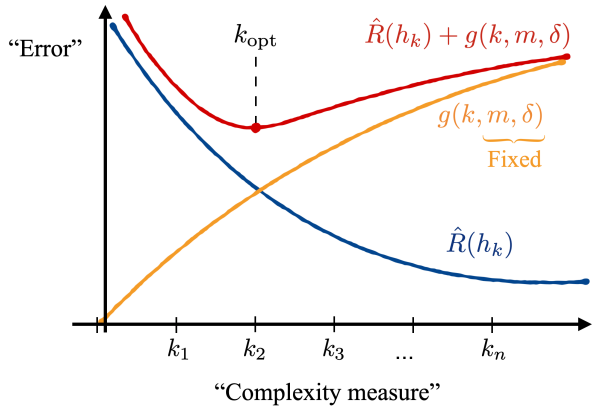


Figure 2: Illustration of structural risk minimization (adapted from Ref. [20]). Increasing the complexity of a hypothesis class typically allows one to obtain hypotheses with decreasing empirical risk. However, in many cases increasing the complexity of a hypothesis class also leads to a larger upper bound on the generalization gap. Structural risk minimization aims to identify a hypothesis with the smallest upper bound on the true risk that quantifies the out-of-sample performance by combining an evaluation of the empirical risk of candidate hypotheses with an upper bound on the generalization gap of the relevant hypothesis class.

which the generalization bound depends explicitly. While increasing k increases the expressivity of the hypothesis class and therefore typically leads to smaller empirical risk, it also increases the upper bound $g(k, m, \delta)$ on the generalization gap and may therefore lead to hypotheses with worse out-of-sample performance. As such, a natural strategy to find an optimal hypothesis – in the sense of having the smallest probabilistic upper bound on the true risk – is as follows:

1. For k in $\{k_1, \ldots, k_n\}$, run the learning algorithm $\mathcal{A}_k$, with hypothesis class $\mathcal{F}_k$, and obtain the hypothesis h_k.
2. Calculate $k_{\text{opt}} = \text{argmin}_k[\hat{R}_S(h_k) + g(k, m, \delta)]$.
3. Output $h_{k_{\text{opt}}}$.

We refer to such a procedure as *structural risk minimization*[2], and contrast this with *empirical risk minimization*, which simply outputs the hypothesis minimizing the empirical risk. In light of the above discussion, we note that, given a family of hypothesis classes $\{\mathcal{F}_k\}$, each specified by some architectural hyper-parameter k and satisfying the condition of Eq. (9), we would ideally like to obtain an upper bound on the generalization gap $g(k, m, \delta)$ which grows *slowly* with respect to k. In particular, we can now understand this from two different but complementary perspectives:

Firstly, from the structural risk minimization (or model selection) perspective, we see from Figure 2 that slow growth of $g(k, m, \delta)$ is indicative of our ability to exploit the expressivity of more complex hypothesis classes, i.e. those with larger k, without risking poor generalization performance due to overfitting. More specifically, under the assumption of monotonically decreasing empirical risk, the slower $g(k, m, \delta)$ grows, the longer we can expect the quantity $\hat{R}_S(h_k) + g(k, m, \delta)$ to decrease before reaching a minimum, and therefore the smaller we can expect our ultimate upper bound on the true risk of the optimal hypothesis $h_{k_{\text{opt}}}$ to be. In contrast, if $g(k, m, \delta)$ grows too fast with respect to k, then even if we can achieve very small empirical risk by increasing model complexity, we do not expect to be able to achieve a sufficiently small upper bound on the true risk of the optimal hypothesis $h_{k_{\text{opt}}}$.

Secondly, from the sample complexity perspective, let us denote by $f(\epsilon, \delta, k)$ the complementary upper bound on the minimum sample sample size m sufficient to probabilistically ensure a generalization gap less than $\epsilon > 0$, which typically follows from $g(k, m, \delta)$ (as we recall from the discussion around Eqs. (7) and (8)). As we naturally expect $g(k, m, \delta)$ to be decreasing with increasing m, slow growth of $g(k, m, \delta)$ with respect to k typically implies slow growth of $f(\epsilon, \delta, k)$

[2]We note that the term "structural risk minimization" is sometimes used to refer to the strategy of minimizing a regularized empirical risk, with an additive regularization term which penalizes high model complexity. However, we follow Ref. [20] in our definition and presentation.

with respect to k. In other words, slow growth of $g(k, m, \delta)$ typically implies slow growth, with respect to model complexity, of the minimum amount of data one has to use before being able to probabilistically guarantee a certain generalization gap for all output hypotheses. As generating data (i.e., sampling from the distribution P) may be expensive or difficult, and as the run-time of learning algorithms typically scales with respect to the data set size, slow growth of $g(k, m, \delta)$ therefore facilitates the process of learning with models of higher complexity.

Given the above observations, we can finally understand the motivation of this work in an informal way. In particular, in the following section we will see that parametrized quantum circuits (PQCs) naturally give rise to hypothesis classes with multiple architectural hyper-parameters, each reflecting a different aspect of the circuit architecture, such as circuit depth, circuit width, the total number of gates or the total number of data-encoding gates of a particular type. In Section 4 we will then see that a body of previous work has resulted in a collection of generalization bounds for PQC-based models, each of which depend explicitly on some subset of architectural hyper-parameters, but not on others. As of yet, however, there exist no generalization bounds which depend explicitly on hyper-parameters associated with the data-encoding strategy, despite the important role such strategies play in determining the expressive power of PQC-based hypothesis classes [33]. As such, the questions which we address in this work are as follows:

(a) Can we derive generalization bounds for PQC-based hypothesis classes which depend explicitly on hyper-parameters associated with the data-encoding strategy?

(b) Can we use such bounds to identify data-encoding strategies for which the upper bounds on the generalization gap grow polynomially with respect to the architectural hyper-parameter relevant to the encoding strategy?

As will be discussed in Section 7, apart from filling a gap in our understanding of the manner in which the data-encoding influences generalization, such bounds would also complement existing works, in that they would allow one to perform structural risk minimization with respect to multiple architectural hyper-parameters simultaneously. With this motivation in mind, before proceeding it is worth briefly mentioning *how* (uniform) generalization bounds are typically obtained. Intuitively, one might expect that the generalization performance of a hypothesis class is related to how *complex* (or how *expressive*) the hypothesis class is, and thus one might hope for the existence of a complexity measure for hypothesis classes from which generalization bounds follow. This intuition is indeed correct, and in fact a large amount of work in statistical learning theory has resulted in a variety of suitable complexity measures – such as the VC dimension [38], Rademacher complexity [39], pseudo-dimension [40] and metric-entropy amongst others – all of which directly give rise to generalization bounds [20, 34, 35]. As a result, given a hypothesis class $\mathcal{F}_k$, one typically proves a uniform generalization bound for $\mathcal{F}_k$, which depends explicitly on the architectural hyper-parameter k, by first characterizing the dependence of a suitable complexity measure C on k (i.e., by writing/bounding $C(\mathcal{F}_k)$ explicitly in terms of k), and then writing down the known generalization bound which follows from $C(\mathcal{F}_k)$. We also follow such a strategy in this work by first characterizing both the Rademacher complexity and metric-entropy of PQC-based models in terms of architectural hyper-parameters related to the data-encoding strategy and then presenting generalization bounds in terms of these complexity measures. At this stage it is hopefully clear, both *why* generalization bounds are desirable, and *how* (at least intuitively) one might obtain such bounds. Given this, we proceed in the following section to define more precisely the PQC-based hypothesis classes considered in this work.

3 Parametrized quantum circuit based model classes

Parametrized quantum circuits (PQCs) are ubiquitous in the field of near-term quantum computing [9–11] and can be used to construct quantum machine learning models [12]. We will consider qubit-based quantum systems. The focus of this work lies on *variational quantum machine learning models* that are constructed from a PQC $U_{\boldsymbol{\theta}}(\boldsymbol{x})$ that depends on trainable parameters $\boldsymbol{\theta} \in \Theta$ and on data inputs $\boldsymbol{x} \in \mathcal{X}$. A prediction in the co-domain $\mathcal{Y} = \mathbb{R}$ is then obtained by evaluating the

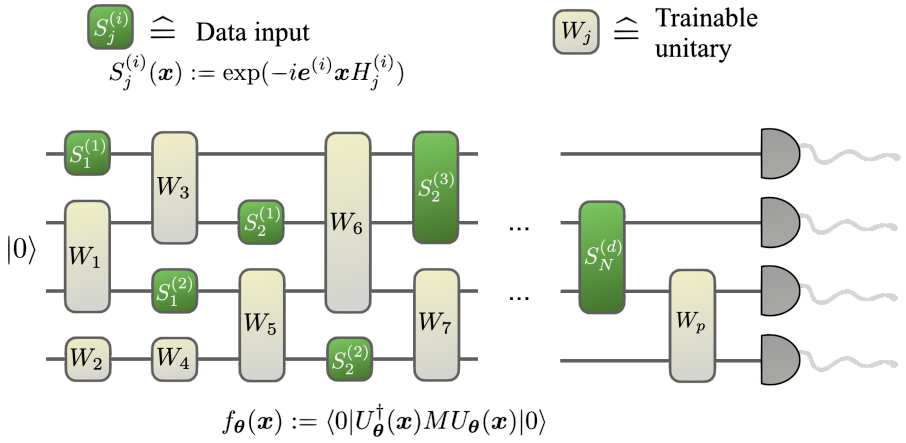


Figure 3: Circuit model considered in this work. We assume that the circuit consists of gates which are parametrized either by the data $\boldsymbol{x}$ (data-encoding gates), or the trainable parameters $\boldsymbol{\theta}$ (trainable gates). The data encoding gates are assumed to implement the time evolution of a data-encoding Hamiltonian, with evolution time given by some data coordinate $x^{(i)} = \boldsymbol{e}^{(i)}\boldsymbol{x}$. The model output is then given by the expectation value of an observable M.

expectation value of a fixed observable M, which can be efficiently evaluated, as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle 0|U_{\boldsymbol{\theta}}^\dagger(\boldsymbol{x})MU_{\boldsymbol{\theta}}(\boldsymbol{x})|0\rangle. \tag{11}$$

In the following, we assume that the data inputs are d-dimensional real-valued vectors with entries in the interval $[0, 2\pi)$, i.e., $\mathcal{X} = [0, 2\pi)^d$. This choice is somewhat arbitrary, as data can always be rescaled to fit into a particular interval. However, $[0, 2\pi)$ is a natural choice because quantum gates available on actual hardware are usually parametrized in terms of *angles*. As will become apparent later, we need not make any assumptions on the nature of the trainable parameters, but in most cases they will also be angles, i.e., $\Theta = [0, 2\pi)^p$, where p is the number of trainable parameters.

We also make some assumptions on the structure of the circuit $U_{\boldsymbol{\theta}}(\boldsymbol{x})$. Our model is motivated by the actual quantum circuits that can be executed on NISQ devices. These devices usually only allow fixed gates and parametrized evolutions under device-specific Hamiltonians [41–43]. In our model, the data inputs $\boldsymbol{x}$ and the trainable parameters $\boldsymbol{\theta}$ enter the circuit through different gates. The unitaries parametrized by $\boldsymbol{\theta}$, denoted by $\{W_i(\boldsymbol{\theta})\}$, constitute the trainable part of the model. Fixed unitaries can be absorbed into the trainable unitaries.

We assume that the gates through which the data enters the circuit are time evolutions under some Hamiltonian, where the "evolution time" is given by one of the data coordinates $x^{(i)}$. We denote the j-th gate that encodes the data coordinate $x^{(i)}$ as

$$S_j^{(i)}(\boldsymbol{x}) = \exp\left(-ix^{(i)}H_j^{(i)}\right) = \exp\left(-i\boldsymbol{e}^{(i)}\boldsymbol{x}H_j^{(i)}\right), \tag{12}$$

where we rewrote the encoding gate in terms of the input data vectors by recognizing that $x^{(i)} = \boldsymbol{e}^{(i)}\boldsymbol{x}$, where $\boldsymbol{e}^{(i)}$ is a standard basis vector. It is of course possible to consider more general dependencies of the evolution time on the input data, i.e. in terms of linear combinations or even non-linear functions of the data coordinates. However, we choose not to include models with such classical pre-processing of the data, in order to isolate the part of the model which is truly quantum. Indeed, if one allowed for arbitrary pre-processing, then one could just use a very complicated neural network to find suitable evolution times for good predictions, but that would miss the point of using a quantum learning model at all. We note though that our definition still encompasses such approaches after a suitable reparametrization of the inputs, which will usually result in a larger number of input coordinates.

For our analysis, no restriction on the placement of the trainable gates and the data-encoding gates in the circuit is necessary. Thus, we assume that they can be arranged arbitrarily, as depicted in Figure 3. However, we will refer to the choice of data-encoding Hamiltonians per data coordinate

as $\mathcal{D}^{(i)} = \{H_j^{(i)}\}$ and call the union of these sets over all data coordinates the *data-encoding strategy*

$$\mathcal{D} = \left(\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(d)}\right). \tag{13}$$

The total number of encoding gates per data coordinate is $N^{(i)} = |\mathcal{D}^{(i)}|$ and the total number of data-encoding gates is $N = \sum_{i=1}^{d} N^{(i)}$.

A data-encoding strategy $\mathcal{D}$ together with a fixed circuit structure and a choice of trainable gates defines a parametrized quantum circuit $U_{\boldsymbol{\theta}}(\boldsymbol{x})$. We denote the fact that this circuit uses the encoding strategy $\mathcal{D}$ as $U_{\boldsymbol{\theta}}(\boldsymbol{x}) \sim \mathcal{D}$. When we fix an observable M to generate the predictions, this defines a function class

$$\mathcal{F}_{\Theta,\mathcal{D},M} \coloneqq \{[0,2\pi)^d \ni \boldsymbol{x} \mapsto \langle 0|U_{\boldsymbol{\theta}}^{\dagger}(\boldsymbol{x}) M U_{\boldsymbol{\theta}}(\boldsymbol{x})|0\rangle \,|\, \boldsymbol{\theta} \in \Theta, U_{\boldsymbol{\theta}}(\boldsymbol{x}) \sim \mathcal{D}\}, \tag{14}$$

which is obtained by considering all possible parametrizations $\boldsymbol{\theta} \in \Theta$ of the trainable gates. This function class depends explicitly on the parametrization of the trainable parts of the circuit and on the data-encoding strategy. As we ultimately want to obtain generalization bounds that depend on the hyper-parameters associated with the encoding strategy – such as the number of encoding gates N – it will be helpful for us to reformulate the function class in a way that makes it more amenable to the analyses in the following sections. To this end, we draw on the results of Refs. [15, 33], which show that the nature of the data encoding gates as Hamiltonian evolutions allows us to expand the model output as a *generalized trigonometric polynomial (GTP)*. A GTP "generalizes" the notion of a trigonometric polynomial by allowing arbitrary frequencies as in

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{\omega} \in \Omega(\mathcal{D})} c_{\boldsymbol{\omega}}(\boldsymbol{\theta}, M) e^{-i\boldsymbol{\omega}\boldsymbol{x}}. \tag{15}$$

While the GTP's coefficients $\{c_{\boldsymbol{\omega}}\}$ depend on the particular parametrization and observable, the set of frequencies $\Omega(\mathcal{D})$ depends solely on the chosen data-encoding strategy $\mathcal{D}$, in particular on the spectra of the Hamiltonians $\{H_j^{(i)}\}$ that yield the data encoding evolutions $\{S_j^{(i)}(\boldsymbol{x})\}$. We describe the procedure for obtaining such a GTP representation in more detail below. The fact that the expectation value is always real is reflected by $c_{\boldsymbol{\omega}} = c_{-\boldsymbol{\omega}}^*$ and by the observation that $\boldsymbol{\omega} \in \Omega(\mathcal{D})$ implies that also $-\boldsymbol{\omega} \in \Omega(\mathcal{D})$. Additionally, we note that the absolute value of any expectation value obtained from measuring M is upper bounded by its operator norm $\|M\|_\infty$, and therefore, if we assume that $\|M\|_\infty \leq B$, then

$$\mathcal{F}_{\Theta,\mathcal{D},M} \subseteq \mathcal{F}_{\Omega}^{B} \coloneqq \left\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{\boldsymbol{\omega} \in \Omega} c_{\boldsymbol{\omega}} \exp(-i\boldsymbol{\omega}\boldsymbol{x}) \,\middle|\, (c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega} \text{ such that } \|f\|_\infty \leq B \right\}, \tag{16}$$

where $\Omega = \Omega(\mathcal{D})$. We have thus defined a function class that solely depends on the data-encoding strategy. We stress that this function class subsumes all possible ways to parametrize the trainable parts of a circuit with fixed data-encoding strategy $\mathcal{D}$ and fixed observable M, but also goes beyond this by allowing all possible choices of observable M such that $\|M\|_\infty \leq B$. Therefore, it also contains models where not only the parameters of the trainable gates, but also the measurement itself is subject to optimization. In going from $\mathcal{F}_{\Theta,\mathcal{D},M}$ to $\mathcal{F}_{\Omega}^{B}$, we effectively allow for a universal trainable part and observable, which enables us to focus on the encoding strategy. Studying intermediate classes between $\mathcal{F}_{\Theta,\mathcal{D},M}$ and $\mathcal{F}_{\Omega}^{B}$ could constitute a path towards tighter generalization bounds that depend on both the data-encoding and the trainable part of the PQC-based model.

In Section 5, we will first prove generalization bounds for $\mathcal{F}_{\Omega}^{B}$, which depend explicitly on properties of Ω, before exploring in detail in Section 6 how these relevant properties of Ω depend on the data-encoding strategy $\mathcal{D}$. Exploiting the fact that, for a given $B \geq \|M\|_\infty$, $\mathcal{F}_{\Theta,\mathcal{D},M} \subseteq \mathcal{F}_{\Omega(\mathcal{D})}^{B}$ then automatically yields explicitly encoding-dependent generalization bounds for $\mathcal{F}_{\Theta,\mathcal{D},M}$.

As the connection between the data-encoding strategy $\mathcal{D}$ and the set $\Omega(\mathcal{D})$ plays a crucial role, we illustrate this connection for a generic data-encoding strategy here. We first consider the action of a single encoding evolution $S(\boldsymbol{x})$ in the density matrix picture, where it acts via the quantum channel

$$\mathcal{S}(\boldsymbol{x})[\rho] = \exp\left(-i\boldsymbol{e}\boldsymbol{x}H\right) \rho \exp\left(i\boldsymbol{e}\boldsymbol{x}H\right), \tag{17}$$

where the Hamiltonian H takes the role of any of the above Hamiltonian terms $H_j^{(i)}$ and $\boldsymbol{e}$ can be any basis vector. We can expand ρ in the eigenbasis of the Hamiltonian $H|\lambda_k\rangle = \lambda_k|\lambda_k\rangle$ and obtain

$$\mathcal{S}(\boldsymbol{x})[\rho] = \mathcal{S}(\boldsymbol{x}) \left[\sum_{k,l} \rho_{k,l} \, |\lambda_k\rangle\langle\lambda_l| \right] \tag{18}$$

$$= \sum_{k,l} \rho_{k,l} \, \mathcal{S}(\boldsymbol{x}) \left[|\lambda_k\rangle\langle\lambda_l| \right] \tag{19}$$

$$= \sum_{k,l} \rho_{k,l} \, \exp(-i(\lambda_k - \lambda_l)\boldsymbol{e}\boldsymbol{x}) |\lambda_k\rangle\langle\lambda_l|. \tag{20}$$

We see that the differences of the eigenvalues λ_k of the Hamiltonian H determine the frequencies with which the different elements of the expansion of ρ are multiplied. We can combine the different frequencies with the weight vector $\boldsymbol{e}$ to obtain the set of all available frequencies

$$\Omega(H) = \{\boldsymbol{\omega}_{k,l} = (\lambda_k - \lambda_l)\boldsymbol{e} \,|\, \lambda_k, \lambda_l \in \mathrm{spec}(H)\}. \tag{21}$$

With this notation, we can simplify our expression for $\mathcal{S}(\boldsymbol{x})[\rho]$ to obtain

$$\mathcal{S}(\boldsymbol{x})[\rho] = \sum_{\boldsymbol{\omega} \in \Omega(H)} \exp(-i\boldsymbol{\omega}\boldsymbol{x}) \rho_{\boldsymbol{\omega}}, \tag{22}$$

where the operators $\rho_{\boldsymbol{\omega}}$ are given by collecting the terms in the above sum for which the frequency differences are the same, i.e.

$$\rho_{\boldsymbol{\omega}} = \sum_{(k,l) \in I(\boldsymbol{\omega})} \rho_{k,l} |\lambda_k\rangle\langle\lambda_l|, \text{ where } I(\boldsymbol{\omega}) = \{(k,l) \,|\, (\lambda_k - \lambda_l)\boldsymbol{e} = \boldsymbol{\omega}\}. \tag{23}$$

As ρ is Hermitian, we have that $\rho_{\boldsymbol{\omega}} = \rho^*_{-\boldsymbol{\omega}}$. The frequency structure carries over if we measure the expectation value of an arbitrary observable M for the state $\mathcal{S}(\boldsymbol{x})[\rho]$ to obtain a prediction

$$f(\boldsymbol{x}) = \mathrm{Tr}\{\mathcal{S}(\boldsymbol{x})[\rho]M\} = \sum_{\boldsymbol{\omega} \in \Omega(H)} \exp(-i\boldsymbol{\omega}\boldsymbol{x}) \, \mathrm{Tr}\{\rho_{\boldsymbol{\omega}} M\} = \sum_{\boldsymbol{\omega} \in \Omega(H)} c_{\boldsymbol{\omega}} \exp(-i\boldsymbol{\omega}\boldsymbol{x}). \tag{24}$$

As a result, we obtain a GTP with coefficients $c_{\boldsymbol{\omega}} = \mathrm{Tr}\{\rho_{\boldsymbol{\omega}} M\}$. Note that, as $\rho_{\boldsymbol{\omega}} = \rho^*_{-\boldsymbol{\omega}}$, we have that $c_{\boldsymbol{\omega}} = c^*_{-\boldsymbol{\omega}}$, which ensures that $f(\boldsymbol{x})$ is real-valued as expected. The coefficients of this series could depend intricately on the circuit that was used to construct ρ and on the specific observable M, but a profound understanding of this relation is an open question. However, this does not pose an obstacle for us, as only the set Ω is relevant for our study.

We have just derived the frequency structure for one encoding gate, but for more complicated circuits we have to understand the action of multiple encoding gates, potentially interleaved with some trainable unitaries. The intermediary unitaries, however, will only result in a basis change, not affecting the set of combined frequencies. We can therefore ignore them and just consider the repeated action of two distinct encoding gates with Hamiltonians H_1 and H_2, resulting in

$$\mathcal{S}_2(\boldsymbol{x})[\mathcal{S}_1(\boldsymbol{x})[\rho]] = \mathcal{S}_2(\boldsymbol{x}) \left[\sum_{\boldsymbol{\omega}_1 \in \Omega(H_1)} \exp(-i\boldsymbol{\omega}_1\boldsymbol{x}) \rho_{\boldsymbol{\omega}_1} \right] \tag{25}$$

$$= \sum_{\boldsymbol{\omega}_1 \in \Omega(H_1)} \exp(-i\boldsymbol{\omega}_1\boldsymbol{x}) \sum_{\boldsymbol{\omega}_2 \in \Omega(H_2)} \exp(-i\boldsymbol{\omega}_2\boldsymbol{x}) \rho_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} \tag{26}$$

$$= \sum_{\boldsymbol{\omega}_1 \in \Omega(H_1)} \sum_{\boldsymbol{\omega}_2 \in \Omega(H_2)} \exp(-i[\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2]\boldsymbol{x}) \rho_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2}. \tag{27}$$

At this point, we precisely understand that the application of the second gate results in new frequencies that encompass all possible sums of the different frequencies. We can again consolidate

this if we consider the sumset (or Minkowski sum) of the two sets of frequencies $\Omega(H_1)$ and $\Omega(H_2)$ defined as

$$\Omega(\{H_1, H_2\}) \coloneqq \Omega(H_1) + \Omega(H_2) \coloneqq \{\boldsymbol{\omega}_1 + \boldsymbol{\omega}_2 \,|\, \boldsymbol{\omega}_1 \in \Omega(H_1), \boldsymbol{\omega}_2 \in \Omega(H_2)\}. \tag{28}$$

With that we have

$$\mathcal{S}_2(\boldsymbol{x})[\mathcal{S}_1(\boldsymbol{x})[\rho]] = \sum_{\boldsymbol{\omega} \in \Omega(H_1)+\Omega(H_2)} \exp(-i\boldsymbol{\omega}\boldsymbol{x})\rho_{\boldsymbol{\omega}}. \tag{29}$$

Note again that the values of specific components $\rho_{\boldsymbol{\omega}}$ depend on the specific initial state ρ and possible intermediate unitaries, but, in this work, we are only interested in Ω itself. We can apply the same logic recursively to see that the set of accessible frequencies for any encoding strategy $\mathcal{D}$ is given by the sumset of all the individual sets of frequencies $\Omega(H_j^{(i)})$ for each gate:

$$\Omega(\mathcal{D}) = \sum_{\mathcal{D}^{(i)} \in \mathcal{D}} \sum_{H \in \mathcal{D}^{(i)}} \Omega(H) = \sum_{i=1}^{d} \sum_{j=1}^{N^{(i)}} \{(\lambda_k - \lambda_l)\boldsymbol{e}^{(i)} \,|\, \lambda_k, \lambda_l \in \mathrm{spec}(H_j^{(i)})\}. \tag{30}$$

4 Prior and related work

Before presenting our explicitly encoding-dependent generalization bounds for PQC-based models in the next two sections, we discuss how our results compare to prior work. While there is a massive amount of prior and ongoing work on the generalization capacity of classical models, see for example the survey in Ref. [37], such results have only recently begun to emerge for PQC-based models. Here, we focus on a comparison with these latter results. Additionally, while the following paragraphs constitute a detailed review of existing generalization bounds for PQC-based models, we stress that no knowledge of these prior works is necessary to understand our proofs and results. In particular, the presentation here is intended to establish context for our work and to place prior works in relation to each other, but the remainder of this manuscript can safely be read independently of the review presented here.

Given the discussions in the previous two sections, we note that, at a high level, all prior work on generalization bounds for PQC-based models can be classified via the following three criteria:

1. Which restrictions – if any – are placed on the architecture/structure of the PQCs generating the model class considered?

2. In terms of which architectural hyper-parameters, or experimentally accessible quantities, are the generalization bounds expressed?

3. Via which complexity measure are the generalization bounds derived?

Given this, we will use the above questions as guidelines for understanding and relating existing results. Throughout this discussion, keep in mind that, as explained in Section 1, all prior works are restricted to *encoding-first* models, whereas we allow for data re-uploading.

Additionally, while some of the following works study the same complexity measures as the ones examined here – namely, Rademacher complexity and covering numbers – all of them differ from ours in both the restriction to encoding-first PQC-based models and in a lack of *explicit* dependence on the data-encoding strategy. Given this, we split our survey into two parts. First, in Section 4.1, we discuss those prior works which derive *encoding-independent* generalization bounds. In Section 4.2, we then discuss existing works deriving generalization bounds which depend on the data-encoding strategy, but with a dependence which is implicit, and not necessarily clear a priori.

4.1 Encoding-independent complexity and generalization bounds

Ref. [24] is an early study of the complexity and generalization capacity of quantum circuit based models, which presents encoding-independent bounds on the pseudo-dimension of function classes associated with encoding-first 2-local (unitary or CPTP) PQCs, polynomial in the size (number of gates) and depth of the trainable part of the circuit (in which all gates were considered trainable).

Such pseudo-dimension bounds then yield generalization bounds, which also depend polynomially on the size and depth of the trainable circuit. Ref. [44] has extended the generalization bounds of Ref. [24] to the agnostic setting. In a similar vein, Ref. [29] has recently derived encoding-independent covering number bounds for encoding-first PQC-based models, which depend explicitly on the number of gates in the PQC, and the operator norm of the measured observable. Once again, using standard tools from statistical learning theory, the authors of Ref. [29] are then able to use these covering number bounds to provide an encoding-independent generalization bound.

Working from the perspective of kernel methods, Ref. [32] has recently investigated the complexity of encoding-first PQC-based models in terms of properties of the parametrized measurement which follows data-encoding. More specifically, they interpret the entire parametrized circuit following the data-encoding as a parametrized measurement, and provide bounds for the VC-dimension of the model class in terms of the rank of the parametrized observable, and for the fat-shattering dimension in terms of the Frobenius norm of the parametrized observable. These bounds on standard complexity measures then allow them to prove generalization bounds which depend explicitly on either the rank or the Frobenius norm of the accessible observables. However, similarly the perspective we advocate in this work, the authors of Ref. [32] stress the application of generalization bounds for model selection, via structural risk minimization.

Finally, Ref. [27] has recently initiated a resource-theoretic approach by providing encoding-independent bounds on both the Rademacher and Gaussian complexity of encoding-first PQC-based models, in terms of the number of repetitions of resource channels allowed in the PQC. These Rademacher and Gaussian complexity bounds have then been used to derive generalization bounds, which depend on the same quantities, and therefore provide an encoding-independent resource-theoretic perspective on generalization in encoding-first PQC-based models.

4.2 Encoding-dependent complexity and generalization bounds

We proceed by discussing prior work deriving generalization bounds which do depend on the data-encoding strategy. While the dependence on the data-encoding could take various forms, in this manuscript we aim to derive generalization bounds which depend *explicitly* on architectural hyper-parameters related to the data-encoding strategy (such as the number of encoding gates of a specific type), and therefore facilitate the straightforward implementation of model selection via structural risk minimization. This is in contrast to all of the prior encoding-dependent generalization bounds, which are written in terms of some quantity which depends on the data-encoding strategy, but with an implicit dependence which is not a priori clear, and needs to be assessed experimentally. Given this fundamental difference between our generalization bounds and those of the prior works we discuss here, a natural open question is whether the implicitly encoding-dependent quantities used in the following works can be written explicitly in terms of architectural hyper-parameters related to the data-encoding strategy. If possible, this would immediately provide explicitly encoding-dependent generalization bounds comparable to those we derive in this work.

With this in mind, we begin our survey of implicitly encoding-dependent generalization bounds with Ref. [25], which has suggested a complexity measure based on the classical Fisher information, called the effective dimension, and demonstrated that one can indeed state generalization bounds in terms of the effective dimension. Utilizing the empirical Fisher information as a tool for approximating the effective dimension, Ref. [25] presented numerical experiments which demonstrate a clear dependence of the effective dimension on the encoding-strategy. However, the explicit dependence of the effective dimension on the encoding strategy is not clear and needs to be evaluated experimentally. Additionally, Ref. [25] also provided a comparison between the effective dimension of PQC-based models and comparable classical models, and demonstrated that PQC-based models can exhibit a higher effective dimension. While not discussed explicitly in Ref. [25], we stress, however, that one should *not* use model complexity (e.g., effective dimension) as the sole criterion for model selection, since model classes with higher effective dimension may have worse generalization behavior than models with a lower effective dimension. Instead, as we advocate in this work, one should ideally use a framework such as structural risk minimization to select a model with the smallest upper bound on out-of-sample performance.

Also working from an information theoretic perspective, and with a focus on the role of data-encoding, Ref. [31] has recently presented generalization bounds for PQC-based models in terms

of information-theoretic quantities describing a notion of mutual information between the post-encoding quantum state $\rho(\boldsymbol{x})$ and the classical data. While these generalization bounds have a strong implicit dependence on the data-encoding strategy, it is once again not immediately clear, apart from in a few special cases, how to explicitly express the suggested complexity measure in terms of architectural hyper-parameters related to the data-encoding strategy.

From a resource theoretic perspective, and complementing Ref. [27], the series of works [26, 28] have further studied the Rademacher complexity of encoding-first PQC-based models. However, unlike in Ref. [27], the Rademacher complexity bounds of Refs. [26, 28] are given in terms of quantities that exhibit an implicit dependence on the data-encoding strategy. More specifically, Ref. [28] provides Rademacher complexity bounds in terms of the size, depth and amount of magic available as a resource. Additionally, Ref. [26] also studies noisy PQC-based models and provides Rademacher complexity bounds in terms of either the Rademacher complexity of the associated noiseless circuit or the free-robustness of the model.

Recently, Ref. [45] has studied generalization for PQC-based models using a hardware efficient ansatz with a specific choice of data-encoding. For this setting, they proved VC-dimension bounds that scale polynomially with the minimum of the number of qubits and the number of trainable layers. In their proofs, they combine light cone arguments with a trigonometric function representation for functions implemented by their ansatz.

Finally, we mention Ref. [30] which has developed techniques for evaluating the potential advantages of quantum kernels over classical kernels. These results are of relevance to this work due to the close relationship between PQC-based models and kernel methods [16]. In a first step, the authors of Ref. [30] suggest the evaluation of a geometric quantity which depends on the chosen quantum feature map and the available training data instances. If the quantum machine learning model passes this first test, a model complexity parameter, which now depends on the quantum encoding and the training data (both instances and labels), should be computed. While these complexity measures can be classically computed in time polynomial in the training data size, analytically determining their exact dependence on the data-encoding can be challenging. This is in contrast to our model complexity bounds, which depend straightforwardly on hyper-parameters associated with the data-encoding strategy, such as the number of encoding gates of a specific type.

5 Generalization bounds for generalized trigonometric polynomials

We recall (from Section 3) that we can prove generalization bounds on $\mathcal{F}_{\Theta,\mathcal{D},M}$, the hypothesis class of interest for a given PQC-based model, by proving generalization bounds on $\mathcal{F}_{\Omega}^{B}$. Recall that $\mathcal{F}_{\Omega}^{B}$ has been defined as the class of generalized trigonometric polynomials (GTPs) with frequencies in Ω and infinity-norm bounded by B as

$$\mathcal{F}_{\Omega}^{B} = \left\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} c_{\boldsymbol{\omega}} \exp(-i\boldsymbol{\omega}\boldsymbol{x}) \;\middle|\; (c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega} \text{ such that } \|f\|_{\infty} \leq B \right\}. \tag{31}$$

In order to prove generalization bounds for $\mathcal{F}_{\Omega}^{B}$, it will be convenient to work with the cosine and sine representation of the complex exponential, and with the norm of the vector of coefficients instead of the norm of the function. Note that, since we have observed in Section 3 that $c_{-\boldsymbol{\omega}} = c_{\boldsymbol{\omega}}^{*}$, we can define, for every $\boldsymbol{\omega} \in \Omega$

$$a_{\boldsymbol{\omega}} \coloneqq c_{\boldsymbol{\omega}} + c_{-\boldsymbol{\omega}} \in \mathbb{R}, \tag{32}$$

$$b_{\boldsymbol{\omega}} \coloneqq \frac{1}{i}(c_{\boldsymbol{\omega}} - c_{-\boldsymbol{\omega}}) \in \mathbb{R}. \tag{33}$$

With these, it further follows that

$$c_{\boldsymbol{\omega}} e^{-i\boldsymbol{\omega}\boldsymbol{x}} + c_{-\boldsymbol{\omega}} e^{i\boldsymbol{\omega}\boldsymbol{x}} = a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x}), \tag{34}$$

which allows us to rewrite the sum in Eq. (31) as a sum of real terms only. If we were only considering frequencies given by real numbers, then it would suffice to sum over the non-negative frequencies in the real sum representation. However, we are dealing with frequency vectors. As this

is the case, we start by removing the zero vector from the set of frequencies to obtain $\Omega^* \coloneqq \Omega \setminus \{0\}$. Note that this is meaningful as $0 \in \Omega$ for any Ω of the form introduced in Section 3. Next, we divide Ω^* into two disjoint parts $\Omega^* = \Omega_+ \cup \Omega_-$, with $\Omega_+ \cap \Omega_- = \emptyset$, such that for every $\boldsymbol{\omega} \in \Omega_+$ we have that $-\boldsymbol{\omega} \in \Omega_-$. We again note that this is possible due to the specific form of the sets Ω discussed in Section 3. In particular, we then have $|\Omega| = 2|\Omega_+| + 1$. Additionally, we make use of a shorthand notation for the vectors $(a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}$ and $(b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}$: We keep the indices outside of the parentheses, but remove the indexing set. Namely we write $(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})$ in place of $(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+})$. We only explicitly write the indexing set at certain points to avoid confusion.

With these notational points in mind,we can rewrite the hypothesis class $\mathcal{F}_\Omega^B$ as

$$\mathcal{F}_\Omega^B = \Bigg\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto f(x) = \frac{a_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} (a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})) \\ \Bigg| \ (a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}) \text{ such that } \|f\|_\infty \leq B \Bigg\}, \tag{35}$$

and we define the class $\mathcal{H}_\Omega^B$ via

$$\mathcal{H}_\Omega^B \coloneqq \Bigg\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto \frac{a_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} (a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})) \\ \Bigg| \ \|(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\|_2 \leq 2(2\pi)^{d/2} B \Bigg\}, \tag{36}$$

where the 2-norm is given by

$$\|(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\|_2 \coloneqq \sqrt{a_0^2 + \sum_{\boldsymbol{\omega}\in\Omega_+} (a_{\boldsymbol{\omega}}^2 + b_{\boldsymbol{\omega}}^2)}. \tag{37}$$

We note that, by construction, $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^B$ holds true. To see this, note that for a function $f \in \mathcal{F}_\Omega^B$ given by $f(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}\in\Omega} \exp(-i\boldsymbol{\omega}\boldsymbol{x}) c_{\boldsymbol{\omega}} = a_0/2 + \sum_{\boldsymbol{\omega}\in\Omega_+} (a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x}))$, we obtain

$$\left\|(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+})\right\|_2 \leq 2\left\|(c_0, (c_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega})\right\|_2 = 2\|f\|_2 \leq 2(2\pi)^{d/2}\|f\|_\infty = 2(2\pi)^{d/2} B. \tag{38}$$

As a consequence of the fact that $\mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^B$, generalization bounds uniform over $\mathcal{H}_\Omega^B$ imply generalization bounds uniform over $\mathcal{F}_\Omega^B$. Therefore, we focus on proving generalization bounds for $\mathcal{H}_\Omega^B$.

Our bounds focus on the dependence of generalization on the frequency spectrum Ω. We obtain these bounds from bounds on the complexity of $\mathcal{H}_\Omega^B$, measured in terms of two complexity measures from classical learning theory, namely the *Rademacher complexity* and the *metric entropy*. We first recall the definitions of these important quantities and then give an overview over our results and proof strategy.

Definition 1 ((Empirical) Rademacher complexity)**.** Let $\mathcal{Z}$ be some data space, $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ a function class, and $S = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) \in \mathcal{Z}^m$. The *empirical Rademacher complexity* of $\mathcal{F}$ with respect to S is defined as

$$\hat{\mathcal{R}}_S(\mathcal{F}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim U(\{-1,1\}^m)} \Big[\sup_{f\in\mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(\boldsymbol{z}_i) \Big], \tag{39}$$

where $U(\{-1,1\}^m)$ denotes the uniform distribution on $\{-1,1\}^m$. The i.i.d. random variables $\sigma_1, \ldots, \sigma_m$ are often called *Rademacher random variables*.

For later use, we note that, if $\mathcal{F} \subseteq \mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$, then, for any $S \in \mathcal{Z}^m$ we have $\hat{\mathcal{R}}_S(\mathcal{F}) \leq \hat{\mathcal{R}}_S(\mathcal{G})$. Next, we introduce our second complexity measure:

Definition 2 (Covering nets, covering number, and metric entropy)**.** Let (X, d) be a (pseudo-)metric space. Let $K \subseteq X$ and let $\varepsilon > 0$. We call $N \subseteq K$ an (interior) ε-covering net of K if for all $x \in K$ there exists a$y \in N$ such that $d(x, y) \leq \varepsilon$. The *covering number* $\mathcal{N}(K, d, \varepsilon)$ is defined as the smallest possible cardinality of an (interior) ε-covering net of K. Finally, we define the *metric entropy* $\log_2 \mathcal{N}(K, d, \varepsilon)$ via a logarithm of the covering number.

For our purposes, the relevant covering numbers are those of $\mathcal{H}_\Omega^B$ with respect to the pseudo-metrics induced by the data-dependent semi-norms $\|\cdot\|_{2,S|_x}$, which, given training data $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, are defined as

$$\|f\|_{2,S|_x} \coloneqq \sqrt{\frac{1}{m}\sum_{i=1}^m |f(\boldsymbol{x}_i)|^2}. \tag{40}$$

In Section 5.1, we prove Rademacher complexity bounds for $\mathcal{H}_\Omega^B$. We do so by understanding $\mathcal{H}_\Omega^B$ as (a subset of) a class of functions implemented by a simple classical *neural network* (NN) with a single hidden layer and with sinusoidal activation functions in the hidden layer. For such NN architectures, we can then apply already known Rademacher complexity bounds. This strategy leads to

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{F}_\Omega^B) \leq \hat{\mathcal{R}}_{S|_x}(\mathcal{H}_\Omega^B) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{|\Omega|}{m}}\right) \tag{41}$$

for a training data set S of size m, with data instances $S|_x = \{x_i\}_{i=1}^m$. Here, the $\tilde{\mathcal{O}}$ refers to the asymptotic behavior as $|\Omega|, m \to \infty$ and hides a logarithmic dependence on $|\Omega|$. (As we are most interested in the dependence on $|\Omega|$, we also hide the dependence on B here.) With these Rademacher complexity bounds at hand, we can then derive generalization guarantees for $\mathcal{H}_\Omega^B$, and thus $\mathcal{F}_\Omega^B$, using a standard generalization bound in terms of the Rademacher complexity. We obtain that for a bounded Lipschitz loss function, with probability $\geq 1-\delta$, the generalization error satisfies

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{|\Omega|}{m}} + \sqrt{\frac{\log(1/\delta)}{m}}\right), \tag{42}$$

uniformly over $f \in \mathcal{H}_\Omega^B$ for training data S of size m. Again, we emphasize the leading-order dependence on $|\Omega|$ and hide other parameters. We note that, without further assumptions, as in classical agnostic learning scenarios, we do not expect a better scaling with respect to m than the Hoeffding-like $\sim 1/\sqrt{m}$.

In Section 5.2, we bound the covering number and metric entropy of $\mathcal{H}_\Omega^B$,and thus of $\mathcal{F}_\Omega^B$.We achieve this by constructing a covering net for $\mathcal{H}_\Omega^B$ from a suitable (finer-grained) covering net of the allowed vectors of Fourier coefficients. Here, we crucially use that $|\Omega|$ determines the dimension of the space in which we have to take these covering nets. With this reasoning, we obtain a metric entropy bound of

$$\log_2 \mathcal{N}(\mathcal{F}_\Omega^B, \|\cdot\|_\infty, \varepsilon) \leq \log_2 \mathcal{N}(\mathcal{H}_\Omega^B, \|\cdot\|_\infty, \frac{\varepsilon}{2}) \leq \tilde{\mathcal{O}}\left(|\Omega| \log(1/\varepsilon)\right), \tag{43}$$

where the $\tilde{\mathcal{O}}$ hides logarithmic dependencies on B and $|\Omega|$. Given these metric entropy bounds, we then use the chaining method to derive empirical Rademacher complexity bounds. Again assuming a bounded Lipschitz loss function, this method yields, with probability $\geq 1-\delta$, a generalization error bound of

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{|\Omega|}{m}} + \sqrt{\frac{\log(1/\delta)}{m}}\right), \tag{44}$$

simultaneously for all $f \in \mathcal{F}_\Omega^B \subseteq \mathcal{H}_\Omega^B$, assuming training data of size m and hiding both logarithmic terms and dependencies on B, the Lipschitz constant, and the bound on the loss. While we see that, with the above definition of $\mathcal{F}_\Omega^B$ and $\mathcal{H}_\Omega^B$, the strategies of Sections 5.1 and 5.2 lead to the same generalization bound in leading order, we nevertheless present both approaches because they yield different results if the assumption on the Fourier coefficients appearing in $\mathcal{F}_\Omega^B$ or $\mathcal{H}_\Omega^B$ is changed from a 2-norm bound to a general p-norm bound.

In the light of the discussion in Section 3, these generalization bounds for classes of generalized trigonometric polynomials imply generalization bounds for PQCs. As we have focused on the dependence on the frequency spectrum in the former, we obtain a focus on the encoding-dependence in the latter. We provide and discuss these results in Section 6.

5.1 Generalization bounds for generalized trigonometric polynomials via Rademacher complexity

We begin our analysis by stating our Rademacher complexity bound for $\mathcal{H}_\Omega^B$. As we will see, this bound is obtained by combining two partial results, and will lead directly to a generalization bound. For ease of notation, we write $K_i \coloneqq \max_{\boldsymbol{\omega}\in\Omega_+}\{|\omega_i|\}$ for $i \in \{1,\dots,d\}$ and $K \coloneqq \sum_i K_i$.

Lemma 3 (Rademacher complexity bounds for GTPs)**.** *Let $d, m \in \mathbb{N}$. Let $S|_x \in (\mathbb{R}^d)^m$. Let $\mathcal{H}_\Omega^B$ be as defined in Eq. (36). The empirical Rademacher complexity of $\mathcal{H}_\Omega^B$ with respect to $S|_x \coloneqq (\boldsymbol{x}_1,\dots,\boldsymbol{x}_m)$ can be upper-bounded as*

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{H}_\Omega^B) \leq \mathcal{O}\left(\frac{\min\left\{\sqrt{\log(2d)}\max\{K,(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega|}\},(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega|\log(|\Omega|)}\right\}}{\sqrt{m}}\right). \tag{45}$$

In order to prove Lemma 3 we state and show two partial results, namely Lemmas 4 and 5. These two Lemmata have slightly different proof strategies, but both are motivated by thinking of generalized trigonometric polynomials as being realized by certain neural network architectures.

Lemma 4 (Empirical Rademacher complexity of $\mathcal{H}_\Omega^B$—Version 1)**.** *Let d, m, $S|_x$, and $\mathcal{H}_\Omega^B$ be as in Lemma 3. Then, the empirical Rademacher complexity of $\mathcal{H}_\Omega^B$ with respect to $S|_x$ can be upper-bounded as*

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{H}_\Omega^B) \leq \mathcal{O}\left(\frac{1}{\sqrt{m}}\max\{K,(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega|}\}\sqrt{\log(2d)}\right). \tag{46}$$

Proof. We prove this statement by constructing a function class that contains $\mathcal{H}_\Omega^B$ and whose empirical Rademacher complexity we are able to upper bound by viewing it as arising from a simple layered neural network (NN) architecture. More specifically, we consider the following class of functions

$$\mathcal{G}_\Omega^B \coloneqq \left\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto \frac{d_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} d_{\boldsymbol{\omega}}\sin(\boldsymbol{\alpha}_{\boldsymbol{\omega}}\boldsymbol{x}+\gamma_{\boldsymbol{\omega}}) \;\middle|\; \|(d_0,(d_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\|_2 \leq 2(2\pi)^{\frac{d}{2}}B,\ \boldsymbol{\alpha}_{\boldsymbol{\omega}} \in \prod_{i=1}^d [-K_i,K_i],\ \gamma_{\boldsymbol{\omega}} \in [-\pi,\pi) \right\}, \tag{47}$$

which can be realized by a NN with a single hidden layer of neurons with sine activation functions, and a linear activation at the output neuron. Here, again $(d_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}$ stands for the vector $(d_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}$. Also, note that for every $\boldsymbol{\omega} \in \Omega_+$, $\boldsymbol{\alpha}_{\boldsymbol{\omega}}$ is a d-dimensional vector and $\gamma_{\boldsymbol{\omega}}$ a real number.

We claim that $\mathcal{H}_\Omega^B \subseteq \mathcal{G}_\Omega^B$. We can prove this inclusion directly by finding the corresponding parameters $(d_0,(d_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})$, $(\gamma_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}$ and $(\boldsymbol{\alpha}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}$ for each element $f \in \mathcal{H}_\Omega^B$, specified by the corresponding $(a_0,(a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}},(b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})$. We can find a valid assignment term by term. We start by noting $d_0 = a_0$. Next, we spell out the term corresponding to the frequency vector $\boldsymbol{\omega}$ with the well-known angle sum trigonometric identity

$$d_{\boldsymbol{\omega}}\sin(\boldsymbol{\alpha}_{\boldsymbol{\omega}}\boldsymbol{x}+\gamma_{\boldsymbol{\omega}}) = d_{\boldsymbol{\omega}}\cos(\gamma_{\boldsymbol{\omega}})\sin(\boldsymbol{\alpha}_{\boldsymbol{\omega}}\boldsymbol{x}) + d_{\boldsymbol{\omega}}\sin(\gamma_{\boldsymbol{\omega}})\cos(\boldsymbol{\alpha}_{\boldsymbol{\omega}}\boldsymbol{x}). \tag{48}$$

Now, for any given $(a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}$ and $(b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}$, we can set

$$d_{\boldsymbol{\omega}} \coloneqq \sqrt{a_{\boldsymbol{\omega}}^2 + b_{\boldsymbol{\omega}}^2},\ \boldsymbol{\alpha}_{\boldsymbol{\omega}} \coloneqq \boldsymbol{\omega},\ \text{and}\ \gamma_{\boldsymbol{\omega}} \coloneqq \arctan(b_{\boldsymbol{\omega}}/a_{\boldsymbol{\omega}}). \tag{49}$$

At this point, it is important to confirm that the assignment is valid within the restrictions imposed in Eq. (47). To begin with, we note that the 2-norm bound from Eq. (38), i.e. $\|(a_0,(a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}},(b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\|_2 \leq$

$2(2\pi)^{\frac{d}{2}}B$, translates directly into $\|(d_0,(d_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\|_2 \leq 2(2\pi)^{\frac{d}{2}}B$, since $d_{\boldsymbol{\omega}}^2 = a_{\boldsymbol{\omega}}^2 + b_{\boldsymbol{\omega}}^2$ for all $\boldsymbol{\omega}$. Additionally, one can also see that the components of $\boldsymbol{\alpha}_{\boldsymbol{\omega}}$ are nothing but the frequencies ω_i for each data coordinate, which fall in the interval $[-K_i, K_i]$ by construction. Finally, as a function arctan can output any angle, choosing the branch $[-\pi, \pi)$ is valid. With these, we reach

$$d_{\boldsymbol{\omega}} \sin(\boldsymbol{\alpha}_{\boldsymbol{\omega}}\boldsymbol{x} + \gamma_{\boldsymbol{\omega}}) = a_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}), \tag{50}$$

which has been our goal.

As $\mathcal{G}_\Omega^B$ arises from a NN whose activation functions are 1-Lipschitz, continuous and antisymmetric, we can use Lemma 16 (stated in the Appendix). For that, we require upper bounds for the 1-norm of the weight vector going into each neuron and for the moduli of the biases. For every neuron in the hidden layer, there are d incoming weights, one for each data dimension, corresponding to the d input neurons. Each component of those weight vectors ($\boldsymbol{\alpha}_{\boldsymbol{\omega}}$ in Eq. (47)) takes values in $\in [-K_i, K_i]$ for some $i \in \{1, \ldots, d\}$, so the 1-norm of such a weight vector is upper bounded by K.

At the output neuron, there are $|\Omega_+|$ incoming weights ($d_{\boldsymbol{\omega}}$ in Eq. (47)) and we have a bound on the 2-norm of this weight vector. Therefore, Hölder's inequality applied to the 2-norm gives the 1-norm bound

$$\|(d_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}\|_1 \leq 2(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega_+|}. \tag{51}$$

With that, we now know that the 1-norm of any weight vector in the NN is upper bounded by $\max\{K, 2(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega_+|}\}$.

Next, we note that the modulus of the biases is at most π in the hidden layer, and $2(2\pi)^{\frac{d}{2}}B$ in the output layer. As a result, we have that the moduli of the biases in the NN are upper bounded by $\max\{\pi, 2(2\pi)^{\frac{d}{2}}B\}$. Now that we have collected all the ingredients, we can plug them into Lemma 16 and obtain the bound

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{G}_\Omega^B) \leq \frac{1}{\sqrt{m}} \left(2\pi \max\{K, 2(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega_+|}\}\sqrt{2\log(2d)} + \max\{\pi, 2(2\pi)^{\frac{d}{2}}B\}\right) \tag{52}$$

$$\leq \mathcal{O}\left(\frac{1}{\sqrt{m}} \max\{K, (2\pi)^{\frac{d}{2}}B\sqrt{|\Omega|}\}\sqrt{\log(2d)}\right), \tag{53}$$

where the $\mathcal{O}$ notation refers to the scaling in $|\Omega|$. As $\mathcal{G}_\Omega^B$ contains $\mathcal{H}_\Omega^B$ as a subset, this bound directly implies

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{H}_\Omega^B) \leq \mathcal{O}\left(\frac{1}{\sqrt{m}} \max\{K, (2\pi)^{\frac{d}{2}}B\sqrt{|\Omega|}\}\sqrt{\log(2d)}\right), \tag{54}$$

which completes the proof. □

In the proof of Lemma 4, we do not bound the empirical Rademacher complexity of $\mathcal{H}_\Omega^B$ directly, rather we embed it into a larger class $\mathcal{G}_\Omega^B$ whose complexity we then bound. However, whereas only a discrete set of frequencies is used in $\mathcal{H}_\Omega^B$, the class $\mathcal{G}_\Omega^B$ allows for a continuum of frequencies. In Lemma 5, we modify the idea of the previous proof to avoid this overcounting of frequencies.

Lemma 5 (Empirical Rademacher complexity of $\mathcal{H}_\Omega^B$—Version 2)**.** *Let d, m, $S|_x$, and $\mathcal{H}_\Omega^B$ be as in Lemma 3. Then, the empirical Rademacher complexity of $\mathcal{H}_\Omega^B$ with respect to $S|_x$ can be upper-bounded as*

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{H}_\Omega^B) \leq \mathcal{O}\left(\frac{(2\pi)^{\frac{d}{2}}B}{\sqrt{m}}\sqrt{|\Omega|\log(|\Omega|)}\right). \tag{55}$$

Proof. Analogously to the proof of Lemma 4, we provide an empirical Rademacher complexity upper bound for a larger function class $\tilde{\mathcal{H}}_\Omega^B$. Along the way, we see that the inclusion $\mathcal{H}_\Omega^B \subseteq \tilde{\mathcal{H}}_\Omega^B$ holds, so that the uniform bound we derive for the larger set is immediately inherited for the smaller one. We start by defining an auxiliary set of functions: let $\mathcal{M}_\Omega$ be the set of generalized trigonometric monomials over $\mathbb{R}^d$ with frequency values in Ω_+, defined as

$$\mathcal{M}_\Omega \coloneqq \{0\} \cup \left\{[0,2\pi)^d \ni \boldsymbol{x} \mapsto \cos(\boldsymbol{\omega}\boldsymbol{x}) \,|\, \boldsymbol{\omega} \in \Omega_+\right\} \cup \left\{[0,2\pi)^d \ni \boldsymbol{x} \mapsto \sin(\boldsymbol{\omega}\boldsymbol{x}) \,|\, \boldsymbol{\omega} \in \Omega_+\right\}. \tag{56}$$

Now, recalling that $|\Omega| = 2|\Omega_+| + 1$, we can define the function class of our current interest as

$$\tilde{\mathcal{H}}_\Omega^B \coloneqq \Bigg\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto b_0 + \Big\langle \boldsymbol{w}, \vec{h}(\boldsymbol{x}) \Big\rangle \\ \Bigg| \; \vec{h} \in (\mathcal{M}_\Omega)^{|\Omega|}, \text{ and } b_0 \in \mathbb{R}, \, \boldsymbol{w} \in \mathbb{R}^{|\Omega|} \text{ such that } \|(b_0, \boldsymbol{w})\|_2 \leq 2(2\pi)^{\frac{d}{2}} B \Bigg\}, \tag{57}$$

where we use the notation $\langle \cdot, \cdot \rangle$ for the standard inner product. Notice how $\tilde{\mathcal{H}}_\Omega^B$ can be seen as a class of functions implemented by a single neuron with identity activation and 2-norm bounded weights, where the input signals have been pre-processed by functions from the specified class $\mathcal{M}_\Omega$. With this, we note the inclusion $\mathcal{H}_\Omega^B \subseteq \tilde{\mathcal{H}}_\Omega^B$.

Next, we use Lemma 15 (stated in the Appendix). To use the result, we note that the activation function of the neuron is the identity $x \mapsto x$ (which is a 1-Lipschitz, anti-symmetric function); that $\mathcal{M}_\Omega$ contains the 0-function; that the modulus of the bias is upper bounded by $2(2\pi)^{\frac{d}{2}} B$; and that we can again use Hölder's inequality applied to the 2-norm to upper bound the 1-norm of the weight vector as $\|(b_0, \boldsymbol{w})\|_1 \leq \sqrt{|\Omega|}\|(b_0, \boldsymbol{w})\|_2 \leq 2(2\pi)^{\frac{d}{2}} B\sqrt{|\Omega|}$. With these, Lemma 15 gives us the upper bound

$$\hat{\mathcal{R}}_{S|_x}(\tilde{\mathcal{H}}_\Omega^B) \leq \frac{2(2\pi)^{\frac{d}{2}} B}{\sqrt{m}} + 2 \cdot 2(2\pi)^{\frac{d}{2}} B\sqrt{|\Omega|} \; \hat{\mathcal{R}}_{S|_x}(\mathcal{M}_\Omega). \tag{58}$$

Hence, in order to proceed we need to find an upper bound for the empirical Rademacher complexity of $\mathcal{M}_\Omega$.

We apply Massart's Lemma (which we recall as Lemma 17 in the Appendix for completeness) for this last step. Let A be the set of generalized trigonometric monomials with frequencies in Ω_+, evaluated on every element of $S|_x = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$, i.e.,

$$A \coloneqq \{(0, \ldots, 0)\} \cup \{(\cos(\boldsymbol{\omega}\boldsymbol{x}_1), \ldots, \cos(\boldsymbol{\omega}\boldsymbol{x}_m)) \mid \boldsymbol{\omega} \in \Omega_+\} \cup \{(\sin(\boldsymbol{\omega}\boldsymbol{x}_1), \ldots, \sin(\boldsymbol{\omega}\boldsymbol{x}_m)) \mid \boldsymbol{\omega} \in \Omega_+\} \subseteq \mathbb{R}^m. \tag{59}$$

Note that, by Hölder's inequality, again applied to the 2-norm, and since sine and cosine take values in $[-1, 1]$, we have that $A \subseteq \mathcal{B}_{\sqrt{m}}(0)$, where $\mathcal{B}_r(\boldsymbol{c})$ is the ball of radius r in 2-norm centered at $\boldsymbol{c}$. Now, we can rewrite the empirical Rademacher complexity and apply Massart's lemma (Lemma 17) to get

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{M}_\Omega) \coloneqq \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{M}_\Omega} \frac{1}{m} \sum_{i=1}^m \sigma_i \, h(\boldsymbol{x}_i) \right] \tag{60}$$

$$= \mathbb{E}_\sigma \left[\sup_{\boldsymbol{a} \in A} \frac{1}{m} \boldsymbol{\sigma}\boldsymbol{a} \right] \tag{61}$$

$$\leq \frac{\sqrt{m}}{m} \sqrt{2\log(|A|)} \tag{62}$$

$$\leq \frac{1}{\sqrt{m}} \sqrt{2\log(|\Omega|)}. \tag{63}$$

Plugging this into Eq. (58), we obtain

$$\hat{\mathcal{R}}_{S|_x}(\tilde{\mathcal{H}}_\Omega^B) \leq \frac{2(2\pi)^{\frac{d}{2}} B}{\sqrt{m}} + 2 \cdot 2(2\pi)^{\frac{d}{2}} B\sqrt{|\Omega|} \cdot \frac{1}{\sqrt{m}} \sqrt{2\log(|\Omega|)} \tag{64}$$

$$\leq \mathcal{O}\left(\frac{(2\pi)^{\frac{d}{2}} B}{\sqrt{m}} \sqrt{|\Omega| \log(|\Omega|)} \right). \tag{65}$$

Recalling again that $\mathcal{H}_\Omega^B \subseteq \tilde{\mathcal{H}}_\Omega^B$ then yields the claimed bound. □

Proof of Lemma 3. This follows directly from combining Lemmas 4 and 5. □

With this Rademacher complexity bound at hand, we can make use of standard tools from classical statistical learning theory to derive a generalization bound.

Theorem 6 (Generalization bound for GTPs—Version 1)**.** *Let $d, m \in \mathbb{N}$. Let $\mathcal{H}^B_\Omega$ be as defined in Eq. (36). Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, c]$ be a bounded loss function such that $\mathbb{R} \ni z \mapsto \ell(y, z)$ is L-Lipschitz for all $y \in \mathbb{R}$. For any $\delta \in (0, 1)$ and for any probability measure P on $[0, 2\pi)^d \times \mathbb{R}$, with probability $\geq 1 - \delta$ over the choice of i.i.d. training data $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \in ([0, 2\pi)^d \times \mathbb{R})^m$ of size m, for every $f \in \mathcal{H}^B_\Omega$, the generalization error can be upper-bounded as*

$$R(f) - \hat{R}_S(f) \leq \mathcal{O}\left(\frac{L \min\left\{ \max\{K, (2\pi)^{\frac{d}{2}} B \sqrt{|\Omega|}\} \sqrt{\log(2d)}, (2\pi)^{\frac{d}{2}} B \sqrt{|\Omega| \log(|\Omega|)} \right\}}{\sqrt{m}} + \frac{\sqrt{\log(1/\delta)}}{\sqrt{m}} \right). \tag{66}$$

Proof. The proof of this theorem consists in combining the standard generalization bound in terms of Rademacher complexity with the Rademacher complexity bounds from Lemma 3. More precisely, we define $\mathcal{G} \subseteq [0, c]^{[0,2\pi)^d \times \mathbb{R}}$ to be the class of functions that can be obtained by post-composing elements of $\mathcal{H}^B_\Omega$ with the loss function ℓ – i.e. we define

$$\mathcal{G} \coloneqq \left\{ [0, 2\pi)^d \times \mathbb{R} \ni (\boldsymbol{x}, y) \mapsto \ell(y, f(\boldsymbol{x})) \mid f \in \mathcal{H}^B_\Omega \right\}. \tag{67}$$

We then have the following generalization bound (see, e.g., Theorem 3.3 in Ref. [20] or Theorem 1.15 in Ref. [35]): For any probability measure P on $[0, 2\pi)^d \times \mathbb{R}$ and for any $\delta > 0$, with probability $\geq 1 - \delta$ over the choice of an i.i.d. training data set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \in ([0, 2\pi)^d \times \mathbb{R})^m$ of size m drawn according to P, we have, for every $g \in \mathcal{G}$,

$$\mathbb{E}_{(\boldsymbol{x},y)\sim P}[g(\boldsymbol{x}, y)] - \frac{1}{m} \sum_{i=1}^m g(\boldsymbol{x}_i, y_i) \leq 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3c \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{68}$$

Note that, when writing $g \in \mathcal{G}$ as $g(\boldsymbol{x}, y) = \ell(y, f(\boldsymbol{x}))$ for some $f \in \mathcal{H}^B_\Omega$, we directly have

$$\mathbb{E}_{(\boldsymbol{x},y)\sim P}[g(\boldsymbol{x}, y)] - \frac{1}{m} \sum_{i=1}^m g(\boldsymbol{x}_i, y_i) = R(f) - \hat{R}_S(f). \tag{69}$$

That is, Eq. (68) indeed provides a high-probability bound on the generalization error. Therefore, we now upper-bound the empirical Rademacher complexity $\hat{\mathcal{R}}_S(\mathcal{G})$. To this end, we use Talagrand's Lemma (going back to Ref. [46]) and our bounds for the empirical Rademacher complexity of $\mathcal{H}^B_\Omega$. As we assume that $\mathbb{R} \ni z \mapsto \ell(y, z)$ is L-Lipschitz for all $y \in \mathbb{R}$, we can apply Talagrand's Lemma (Lemma 18) and Lemma 3 to obtain

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{G}) &= \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^m \sigma_i g(x_i, y_i) \right] && (70)\\
&= \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}^B_\Omega} \sum_{i=1}^m \sigma_i \ell(y_i, f(x_i)) \right] && (71)\\
&\leq \frac{L}{m} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}^B_\Omega} \sum_{i=1}^m \sigma_i f(x_i) \right] && (72)\\
&= L \hat{\mathcal{R}}_{S|_x}(\mathcal{H}^B_\Omega) && (73)\\
&\leq \mathcal{O}\left(L \frac{\min\left\{ \sqrt{\log(2d)} \max\{K, (2\pi)^{\frac{d}{2}} B \sqrt{|\Omega|}\}, (2\pi)^{\frac{d}{2}} B \sqrt{|\Omega| \log(|\Omega|)} \right\}}{\sqrt{m}} \right), && (74)
\end{aligned}$$

where we have denoted by $S|_x \coloneqq \{\boldsymbol{x}_i\}_{i=1}^m$ the set of unlabeled training data points. Inserting this bound into Eq. (68) now gives the stated generalization error bound. □

The generalization bound of Theorem 6 can be rewritten as an upper bound on the number of labeled training examples that suffice to guarantee small generalization error.

Corollary 7 (Number of labeled training examples sufficient for a small generalization error—Version 1)**.** *For any $\varepsilon, \delta \in (0,1)$ and for any probability measure P on $[0,2\pi)^d \times \mathbb{R}$, a training data size*

$$m = m(\varepsilon,\delta) \leq \mathcal{O}\left(\frac{L^2 \min\left\{\max\{K^2, (2\pi)^d B^2 |\Omega|\}\log(2d), (2\pi)^d B^2 |\Omega| \log(|\Omega|)\right\}}{\varepsilon^2} + \frac{c^2 \log(1/\delta)}{\varepsilon^2}\right) \tag{75}$$

suffices to guarantee that, with probability $\geq 1-\delta$ over the choice of i.i.d. training data $S \in ([0,2\pi)^d \times \mathbb{R})^m$ of size m, $R(f) - \hat{R}_S(f) \leq \varepsilon$ holds for every $f \in \mathcal{H}_\Omega^B$.

Proof. We set the upper bound on the generalization error proven in Theorem 6 equal to ε and solving for m. □

Remark 8. The proof strategy for obtaining Rademacher complexity bounds of generalized trigonometric polynomials presented here easily extends beyond the case in which the 2-norm of the vector of Fourier coefficients is assumed to be bounded. Namely, if we consider, for $1 \leq p \leq \infty$, the class

$$\mathcal{H}_\Omega^{\tilde{B},p} := \left\{ [0,2\pi)^d \ni \boldsymbol{x} \mapsto \frac{a_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} \left(a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})\right) \;\middle|\; \|(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\|_p \leq \tilde{B} \right\}, \tag{76}$$

with Fourier coefficients of a bounded p-norm, we obtain, with essentially the same proof, an empirical Rademacher complexity bound of

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{H}_\Omega^{\tilde{B},p}) \leq \tilde{\mathcal{O}}\left(\frac{\tilde{B}|\Omega|^{\frac{1}{q}}}{\sqrt{m}}\right), \tag{77}$$

where $q \in [0,1]$ is the Hölder conjugate of p, i.e., $1/p + 1/q = 1$, and the $\tilde{\mathcal{O}}$ hides a logarithmic dependence on $|\Omega|$. This, in turn, leads (for c-bounded L-Lipschitz loss) to a generalization error bound of

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\frac{L\tilde{B}|\Omega|^{\frac{1}{q}} + c\sqrt{\log(1/\delta)}}{\sqrt{m}}\right), \tag{78}$$

which holds with probability $\geq 1-\delta$ uniformly over $\mathcal{H}_\Omega^{B,p}$, for training data of size m. These bounds based on p-norms might be of independent interest. For example, depending on the structure of the trainable part of the PQC, a detailed analysis might lead to additional structural properties (such as sparsity) of the set of admissible Fourier coefficients, which could then lend themselves to an analysis in terms of p-norms for $p \neq 2$.

5.2 Generalization bounds for generalized trigonometric polynomials via covering numbers

Similarly to Section 5.1, we first prove a bound on a complexity measure for the hypothesisclass $\mathcal{F}_\Omega^B$ and then derive a generalization bound from it. This subsection differs from the previous one in that we discuss a different complexity measure, covering numbers, and that we do not need to resort to the larger hypothesis class $\mathcal{H}_\Omega^B$, but rather study $\mathcal{F}_\Omega^B$ directly.

Lemma 9 (Covering number bound for GTPs)**.** *Let $d \in \mathbb{N}$ and $\varepsilon > 0$. Let $\mathcal{F}_\Omega^B$ be as defined in Eq. (16). The ε-covering number of $\mathcal{F}_\Omega^B$ with respect to $\|\cdot\|_\infty$ can be upper-bounded as*

$$\mathcal{N}(\mathcal{F}_\Omega^B, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}(\mathcal{H}_\Omega^B, \|\cdot\|_\infty, \varepsilon/2) \leq \left(\frac{2\cdot 3\cdot 2(2\pi)^{\frac{d}{2}} B\sqrt{|\Omega|}}{\varepsilon}\right)^{|\Omega|}. \tag{79}$$

Therefore, the corresponding metric entropy can be upper-bounded as

$$\log_2 \mathcal{N}(\mathcal{F}_\Omega^B, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{O}\left(|\Omega|[\log((2\pi)^{\frac{d}{2}} B) + \log(|\Omega|) + \log(1/\varepsilon)]\right). \tag{80}$$

Proof. As discussed after introducing the class $\mathcal{H}^B_\Omega$, we have $\mathcal{F}^B_\Omega \subseteq \mathcal{H}^B_\Omega$. Therefore, according to the approximate monotonicity of covering numbers (see, e.g., Exercise 4.2.10 in [47]), we have, for every $\varepsilon > 0$,

$$\mathcal{N}(\mathcal{F}^B_\Omega, \|\cdot\|_\infty, \varepsilon) \leq \mathcal{N}(\mathcal{H}^B_\Omega, \|\cdot\|_\infty, \varepsilon/2). \tag{81}$$

Thus, it remains to prove a covering number bound for $\mathcal{H}^B_\Omega$.

Let $\mathcal{N}_{\tilde{\varepsilon}}$ be an $\tilde{\varepsilon}$-covering net of the ball

$$\mathcal{B} \coloneqq \left\{ \boldsymbol{\xi} = (a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}) \in \mathbb{R}^{|\Omega|} \;\middle|\; \|\boldsymbol{\xi}\|_2 \leq 2(2\pi)^{\frac{d}{2}} B \right\} \tag{82}$$

with respect to the metric induced by $\|\cdot\|_2$ on $\mathbb{R}^{|\Omega|}$. By definition of $\mathcal{H}^B_\Omega$, to every $f \in \mathcal{H}^B_\Omega$ we can associate a point $(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}) \in \mathcal{B}$ such that

$$f(x) = \frac{a_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} \left(a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})\right). \tag{83}$$

Given such a vector of coefficients $(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}) \in \mathcal{B}$ – which, again, for the sake of notational ease, we write as $(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}))$, omitting the Ω_+ everywhere – we can find an element $(\tilde{a}_0, (\tilde{a}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}, (\tilde{b}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\Omega_+}) \in \mathcal{N}_{\tilde{\varepsilon}}$ of the cover that is $\tilde{\varepsilon}$ close in 2-norm to the coefficients of f, i.e., such that

$$\left\|(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}) - (\tilde{a}_0, (\tilde{a}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (\tilde{b}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\right\|_2 \leq \tilde{\varepsilon}. \tag{84}$$

Define $\tilde{f}$ as the function specified by these new coefficients,

$$\tilde{f}(x) = \frac{\tilde{a}_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} \left(\tilde{a}_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + \tilde{b}_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})\right). \tag{85}$$

We now bound the infinity norm distance between f and $\tilde{f}$ in terms of the 2-norm distance between the corresponding coefficients as

$$\begin{aligned}
\left\|f - \tilde{f}\right\|_\infty &\coloneqq \sup_{\boldsymbol{x}\in[0,2\pi)} \left|f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})\right| && (86)\\
&\leq \left|\frac{a_0}{2} - \frac{\tilde{a}_0}{2}\right| + \sup_{\boldsymbol{x}} \sum_{\boldsymbol{\omega}\in\Omega_+} \left|(a_{\boldsymbol{\omega}} - \tilde{a}_{\boldsymbol{\omega}}) \cos(\boldsymbol{\omega}\boldsymbol{x}) + (b_{\boldsymbol{\omega}} - \tilde{b}_{\boldsymbol{\omega}}) \sin(\boldsymbol{\omega}\boldsymbol{x})\right| && (87)\\
&\leq |a_0 - \tilde{a}_0| + \sum_{\boldsymbol{\omega}\in\Omega_+} \left(|a_{\boldsymbol{\omega}} - \tilde{a}_{\boldsymbol{\omega}}| + |b_{\boldsymbol{\omega}} - \tilde{b}_{\boldsymbol{\omega}}|\right) && (88)\\
&= \left\|(a_0, (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (b_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}) - (\tilde{a}_0, (\tilde{a}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (\tilde{b}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}})\right\|_1 && (89)\\
&\leq \sqrt{|\Omega|}\, \tilde{\varepsilon}. && (90)
\end{aligned}$$

Here, we have used the triangle inequality and the fact that sine and cosine can only take values in $[-1, 1]$, as well as (in the last step) Hölder's inequality with respect to the 2-norm. That means, if we denote by $\mathcal{N}_{\mathcal{F}}$ the set of GTPs whose coefficients come from the cover $\mathcal{N}_{\tilde{\varepsilon}}$, i.e.,

$$\mathcal{N}_{\mathcal{F}} \coloneqq \left\{ [0, 2\pi)^d \ni \boldsymbol{x} \mapsto \frac{\tilde{a}_0}{2} + \sum_{\boldsymbol{\omega}\in\Omega_+} \left(a_{\boldsymbol{\omega}} \cos(\boldsymbol{\omega}\boldsymbol{x}) + b_{\boldsymbol{\omega}} \sin(\boldsymbol{\omega}\boldsymbol{x})\right) \;\middle|\; (\tilde{a}_0, (\tilde{a}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}, (\tilde{b}_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}) \in \mathcal{N}_{\tilde{\varepsilon}} \right\}, \tag{91}$$

and if we fix $\tilde{\varepsilon}$ to be $\tilde{\varepsilon} = \varepsilon/\sqrt{|\Omega|}$, then $\mathcal{N}_{\mathcal{F}}$ is an ε-covering net of $\mathcal{H}^B_\Omega$ with respect to $\|\cdot\|_\infty$. Thus, to finish the proof, it remains to upper bound the cardinality $|\mathcal{N}_{\mathcal{F}}| \leq |\mathcal{N}_{\tilde{\varepsilon}}|$. To obtain such a bound, we recall that we only require $\mathcal{N}_{\tilde{\varepsilon}}$ to be an $\tilde{\varepsilon}$-cover of a 2-norm ball of radius $2(2\pi)^{\frac{d}{2}} B$ in $\mathbb{R}^{|\Omega|}$ with respect to the 2-norm. A simple volumetric argument (presented, e.g., in section 4 of Ref. [47]) shows that there exists such a $\tilde{\varepsilon}$-cover $\mathcal{N}_{\tilde{\varepsilon}}$ of $\mathcal{B}$ with cardinality

$$|\mathcal{N}_{\tilde{\varepsilon}}| \leq \left(\frac{3 \cdot 2(2\pi)^{\frac{d}{2}} B}{\tilde{\varepsilon}}\right)^{|\Omega|} = \left(\frac{3 \cdot 2(2\pi)^{\frac{d}{2}} B \sqrt{|\Omega|}}{\varepsilon}\right)^{|\Omega|}. \tag{92}$$

All in all, we have proven that there exists an ε-covering net of $\mathcal{H}_\Omega^B$ with respect to $\|\cdot\|_\infty$ whose cardinality is bounded by

$$\left(\frac{3\cdot 2(2\pi)^{\frac{d}{2}}B\sqrt{|\Omega|}}{\varepsilon}\right)^{|\Omega|}. \tag{93}$$

This is exactly the claimed upper bound on the ε-covering number of $\mathcal{H}_\Omega^B$, thus completing the proof. □

The covering number bound just established implies a generalization bound for GTPs.

Theorem 10 (Generalization bound for generalized trigonometric polynomials—Version 2)**.** *Let $d, m \in \mathbb{N}$. Let $\mathcal{F}_\Omega^B$ be as defined in Eq. (16). Let $\ell : \mathbb{R}\times\mathbb{R} \to [0,c]$ be a bounded loss function such that $\mathbb{R} \ni z \mapsto \ell(y,z)$ is L-Lipschitz for all $y \in \mathbb{R}$. For any $\delta \in (0,1)$ and for any probability measure P on $[0,2\pi)^d \times \mathbb{R}$, with probability $\geq 1-\delta$ over the choice of i.i.d. training data $S \in ([0,2\pi)^d\times\mathbb{R})^m$ of size m, for every $f \in \mathcal{F}_\Omega^B$, the generalization error can be upper-bounded as*

$$R(f) - \hat{R}_S(f) \leq \mathcal{O}\left(BL\sqrt{\frac{|\Omega|(\log(|\Omega|)+\log((2\pi)^{\frac{d}{2}}B))}{m}} + c\sqrt{\frac{\log(1/\delta)}{m}}\right) \tag{94}$$

Proof. The proof consists of three steps. First, we use the chaining technique from random process theory to upper bound the (empirical) Rademacher complexity in terms of an integral over the square root of the uniform empirical metric entropy. Second, we show that the metric entropy with respect to $\|\cdot\|_\infty$ upper-bounds the uniform empirical metric entropy, so we can use the bound in Lemma 9 to upper-bound the (empirical) Rademacher complexity of generalized trigonometric polynomials. Third, we again use the standard generalization bound based on empirical Rademacher complexities.

Similarly to the proof of Theorem 6, we define

$$\mathcal{G} \coloneqq \left\{[0,2\pi)^d \times \mathbb{R} \ni (\boldsymbol{x}, y) \mapsto \ell(y, f(\boldsymbol{x})) \;\middle|\; f \in \mathcal{F}_\Omega^B\right\}. \tag{95}$$

Again, since we assume that $\mathbb{R} \ni z \mapsto \ell(y,z)$ is L-Lipschitz for all $y \in \mathbb{R}$, Talagrand's Lemma (Lemma 18 in the Appendix) tells us that

$$\hat{\mathcal{R}}_S(\mathcal{G}) \leq L\hat{\mathcal{R}}_{S|_x}(\mathcal{F}_\Omega^B), \tag{96}$$

where we have denoted by $S|_x \coloneqq \{\boldsymbol{x}_i\}_{i=1}^m$ the unlabeled training data points. Next, Dudley's Theorem (which we recall as Theorem 19 in the Appendix), yields

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{F}_\Omega^B) \leq \frac{12}{\sqrt{m}}\int_0^{\gamma_0}\sqrt{\log \mathcal{N}(\mathcal{F}_\Omega^B, \|\cdot\|_{2,S|_x}, \beta)}\,\mathrm{d}\beta, \tag{97}$$

where $\|\cdot\|_{2,S|_x}$ is the (data-dependent) semi-norm on $\mathbb{R}^{\mathbb{R}^d}$ defined as $\|f\|_{2,S|_x} \coloneqq \left(\frac{1}{m}\sum_{i=1}^m |f(\boldsymbol{x}_i)|^2\right)^{1/2}$, and we have defined $\gamma_0 \coloneqq \sup_{f\in\mathcal{F}_\Omega^B}\|f\|_{2,S}$.

Now, we note that, for every $f \in \mathcal{F}_\Omega^B$, $\|f\|_{2,S|_x} \leq \|f\|_\infty$. Therefore, we have both that $\gamma_0 \leq \sup_{f\in\mathcal{F}_\Omega^B}\|f\|_\infty \leq B$ and, that for every $\beta > 0$, $\mathcal{N}(\mathcal{F}_\Omega^B, \|\cdot\|_{2,S|_x}, \beta) \leq \mathcal{N}(\mathcal{F}_\Omega^B, \|\cdot\|_\infty, \beta)$. Hence,

we can combine Eq. (97) with our covering number bound from Lemma 9 and further upper bound

$$\hat{\mathcal{R}}_{S|_x}(\mathcal{F}_\Omega^B) \leq \frac{12}{\sqrt{m}} \int_0^{\gamma_0} \sqrt{|\Omega| \left(\log(3 \cdot 2(2\pi)^{\frac{d}{2}} B) + \log(\sqrt{|\Omega|}) + \log\left(\frac{2}{\beta}\right) \right)} \, \mathrm{d}\beta \tag{98}$$

$$\leq \frac{12}{\sqrt{m}} \sqrt{|\Omega|} \left(\gamma_0 \sqrt{\log(3 \cdot 2(2\pi)^{\frac{d}{2}} B) + \frac{1}{2}\log(|\Omega|)} + \int_0^{\gamma_0} \sqrt{\log\left(\frac{2}{\beta}\right)} \, \mathrm{d}\beta \right) \tag{99}$$

$$\leq \frac{12}{\sqrt{m}} \sqrt{|\Omega|} \Bigg(B \sqrt{\log(3 \cdot 2(2\pi)^{\frac{d}{2}} B) + \frac{1}{2}\log(|\Omega|)} \tag{100}$$

$$+ B \sqrt{\log\left(\frac{1}{2(2\pi)^{\frac{d}{2}} B}\right)} - \frac{\sqrt{\pi}}{2} \operatorname{erf}\left(\sqrt{\log\left(\frac{1}{2(2\pi)^{\frac{d}{2}} B}\right)} \right) \Bigg) \tag{101}$$

$$\leq \mathcal{O}\left(B \sqrt{\frac{|\Omega|(\log((2\pi)^{\frac{d}{2}} B) + \log(|\Omega|))}{m}} \right), \tag{102}$$

where we have used the integral

$$\int \sqrt{\log 1/x} \, \mathrm{d}x = x\sqrt{\log 1/x} - (\sqrt{\pi}/2) \cdot \operatorname{erf}(\sqrt{\log 1/x}), \tag{103}$$

with the error function defined as

$$\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, \mathrm{d}t. \tag{104}$$

At this point, we again have a bound on the empirical Rademacher complexity at our disposal. So, just like in the proof of Theorem 6, we can now apply the standard Rademacher complexity generalization bound. This then tells us that, for any probability measure P on $[0, 2\pi)^d \times \mathbb{R}$ and for any $\delta > 0$, with probability $\geq 1 - \delta$ over the choice of an i.i.d. training data set S of size m, we have, for every $f \in \mathcal{F}_\Omega^B$,

$$R(f) - \hat{R}_S(f) \leq 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3c\sqrt{\frac{\log(2/\delta)}{2m}} \tag{105}$$

$$\leq \mathcal{O}\left(BL \sqrt{\frac{|\Omega|(\log(|\Omega|) + \log((2\pi)^{\frac{d}{2}} B))}{m}} + c\sqrt{\frac{\log(1/\delta)}{m}} \right), \tag{106}$$

as claimed. □

Also for this generalization bound, we provide the reformulation in terms of a bound on the sample size sufficient to guarantee small generalization error.

Corollary 11 (Number of labeled training examples sufficient for a small generalization error—Version 2)**.** *Let $d \in \mathbb{N}$. Let $\mathcal{F}_\Omega^B$ Eq. (16).Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, c]$ be an L-Lipschitz loss function. For any $\varepsilon, \delta \in (0, 1)$ and for any probability measure P on $[0, 2\pi)^d \times \mathbb{R}$, a training data size*

$$m = m(\varepsilon, \delta) \leq \mathcal{O}\left(B^2 L^2 \frac{|\Omega|(\log(|\Omega|) + \log((2\pi)^{\frac{d}{2}} B))}{\varepsilon^2} + c^2 \frac{\log(1/\delta)}{\varepsilon^2} \right), \tag{107}$$

suffices to guarantee that, with probability $\geq 1 - \delta$ over the choice of i.i.d. training data $S \in (\mathbb{R}^d \times \mathbb{R})^m$ of size m, $R(f) - \hat{R}_S(f) \leq \varepsilon$ holds for every $f \in \mathcal{F}_\Omega^B$.

Proof. We set the upper bound on the generalization error proven in Theorem 10 equal to ε and solving for m. □

Remark 12. Our metric entropy bounds of trigonometric polynomials presented here again extend beyond the case of bounded 2-norm of the vector of Fourier coefficients to a general bounded p-norm. However, if we again consider, for $1 \leq p \leq \infty$, the class $\mathcal{H}_{\Omega}^{B,p}$ defined in Remark 8, our proof strategy here yields essentially – i.e., to leading order in $|\Omega|$ – the same metric entropy and generalization bounds as for $p = 2$. The reason is that the dimension of the space in which we take covering nets in the proof of Lemma 9 remains proportional to $|\Omega|$, independently of p. We only see improvements for $1 \leq p < 2$ in the terms depending logarithmically on $|\Omega|$. Therefore, while the proof strategies of Sections 5.1 and 5.2 give essentially the same generalization guarantees for $p = 2$, the approach of Section 5.1 adapts nicely to the case $p < 2$, whereas the reasoning of Section 5.2 is typically preferable for $p > 2$.

Remark 13. The proof of Theorem 10 extends beyond Lipschitz loss functions. For example, suppose that $\mathbb{R} \ni z \mapsto \ell(y, z)$ is α-Hölder continuous with Hölder coefficient $A > 0$ for all $y \in \mathbb{R}$, where $\alpha \in (0, 1)$. Then, with the notation of the above proof,

$$\mathcal{N}(\mathcal{G}, \|\cdot\|_{2,S}, \beta) \leq \mathcal{N}\left(\mathcal{F}_{\Omega}^{B}, \|\cdot\|_{2\alpha,S|_x}, (\beta/A)^{1/\alpha}\right). \tag{108}$$

We can thus apply Dudley's Theorem to upper bound

$$\hat{\mathcal{R}}_S(\mathcal{G}) \leq \frac{12}{\sqrt{m}} \int_0^{\tilde{\gamma}_0} \sqrt{\mathcal{N}\left(\mathcal{F}_{\Omega}^{B}, \|\cdot\|_{2\alpha,S|_x}, (\beta/A)^{1/\alpha}\right)}\, \mathrm{d}\beta. \tag{109}$$

Now, we again observe that $\|\cdot\|_{2\alpha,S|_x} \leq \|\cdot\|_{\infty}$ and upper bound the covering number integral, using our result from Lemma 9. The parameters of the Hölder continuity enter the final Rademacher complexity bound via a term scaling with $\sqrt{\log(A)/\alpha}$.

6 Encoding-dependent generalization bounds for parametrized quantum circuits

We are finally in a position to answer the questions posed in Section 2. Recall that our first goal was to derive generalization bounds for PQC-based models which depend explicitly on architectural hyper-parameters related to the data-encoding strategy. We showed in Section 3 how PQC-based model classes can be viewed as a subset of generalized trigonometric polynomials (GTPs), whose set of frequencies Ω is determined solely by the data-encoding strategy $\mathcal{D}$. We then derived complexity and generalization bounds for GTPs in terms of the number of different frequencies $|\Omega(\mathcal{D})|$. In order to provide explicitly encoding-dependent generalization bounds for PQC-based models, it remains to express $|\Omega(\mathcal{D})|$ in terms of the relevant architectural hyper-parameters associated with different data-encoding strategies.

To do so, we recall that the data-encoding strategy of a PQC-based model class is defined as a collection of lists of data-encoding Hamiltonians $\mathcal{D}^{(i)} = \{H_j^{(i)}\}$ associated with each coordinate $x^{(i)}$. We distinguish different data-encoding strategies according to the different assumptions made on the structure of the data-encoding Hamiltonians $H \in \mathcal{D}^{(i)}$. Given a particular assumption, for example that all H are tensor products of Pauli operators or at most κ-local, the natural hyper-parameter associated with the data encoding strategy is the number $N = \sum_{i=1}^{d} |\mathcal{D}^{(i)}|$ of data-encoding Hamiltonians of the assumed type. Hence, our goal in this section is to derive, for different data-encoding strategies, upper bounds on $|\Omega(\mathcal{D})|$ that depend on N as well as as on other relevant properties of the data-encoding Hamiltonians (such as, e.g., the locality κ). By substituting these upper bounds on $|\Omega|$ into the GTP generalization bounds of the previous section, we then obtain generalization bounds for PQC-based model classes which depend explicitly on properties of the data-encoding strategy.

We first recall the definition of Ω from Eq. (30). If we denote the Hamiltonians of the data-encoding strategy associated with $x^{(i)}$ as $\{H_j^{(i)}\}$, we can group the frequencies associated with each data coordinate into a separate sumset $\Omega^{(i)}$:

$$\Omega(\mathcal{D}) = \sum_{i=1}^{d} \sum_{j=1}^{N^{(i)}} \Omega\left(H_j^{(i)}\right) = \sum_{i=1}^{d} \Omega^{(i)}. \tag{110}$$

The frequencies belonging to the different coordinates $\{x^{(i)}\}$ are linearly independent because they were defined to be multiples of different standard basis vectors $\boldsymbol{e}^{(i)}$. This implies that the cardinality of the full set is equal to the product of the individual cardinalities,

$$|\Omega| = \prod_{i=1}^{d} |\Omega^{(i)}|, \tag{111}$$

thus allowing us to multiply bounds on the cardinalities obtained for the separate data-encoding strategies, $|\Omega^{(i)}|$, to obtain a bound on $|\Omega|$.

As the underlying frequencies in $\Omega^{(i)}$ are all scalar multiples of the same basis vector $\boldsymbol{e}^{(i)}$, the analysis of $\Omega^{(i)}$ comes down to the different frequencies generated by the Hamiltonians that are used to encode $x^{(i)}$. For a given single Hamiltonian H, we denote this set by

$$\Delta(H) \coloneqq \{\lambda_i - \lambda_j \,|\, \lambda_i, \lambda_j \in \mathrm{spec}(H)\} \tag{112}$$

so that

$$\Omega(H) = \{\, \delta \boldsymbol{e} \,|\, \delta \in \Delta\,(H)\}\,, \tag{113}$$

where $\boldsymbol{e}$ is the basis vector associated to the respective coordinate. Next, we derive some bounds on $|\Omega^{(i)}|$ for different assumptions on the underlying Hamiltonians.

Worst case upper bounds. We first derive the worst-case limits of $|\Omega^{(i)}|$ for κ-local encoding Hamiltonians. A κ-local Hamiltonian H has local dimension 2^κ and the number of possible differences of eigenvalues in the spectrum is thus upper bounded as

$$|\Delta(H)| \leq \frac{2^\kappa(2^\kappa - 1)}{2} + 1 = \mathcal{O}(2^{2\kappa}). \tag{114}$$

One can in principle construct a Hamiltonian that saturates this bound by choosing $\mathrm{spec}(H_{\max}) = \{0, 3, 9, \ldots, 3^{2^\kappa}\}$, but this is a rather synthetic example that we do not expect to encounter on real hardware. Eq. (114) implies that repeating $N^{(i)}$ κ-local Hamiltonians will, in the case where there are no duplicates in the frequency set, imply a cardinality of at most

$$|\Omega^{(i)}| \leq \left(\frac{2^\kappa(2^\kappa - 1)}{2} + 1\right)^{N^{(i)}} = \mathcal{O}(2^{2\kappa N^{(i)}}). \tag{115}$$

Again, this bound can be saturated by choosing Hamiltonians with ever-larger spectra, namely by choosing $H_1^{(i)} = H_{\max}$ and $\mathrm{spec}(H_{j+1}^{(i)}) = \max(\mathrm{spec}(H_j^{(i)})) \cdot \mathrm{spec}(H_{\max})$.

Repeated Hamiltonians. We now consider the case where the same Hamiltonian $H^{(i)}$ is used $N^{(i)}$ times to encode the coordinate $x^{(i)}$. Due to the underlying symmetry of the definition of $\Delta(H^{(i)})$, we have that

$$\Delta(H^{(i)}) = \{0, \pm\delta_1, \ldots, \pm\delta_T\} \tag{116}$$

for some T, and therefore $|\Delta(H^{(i)})| = 2T+1 = |\Omega_j^{(i)}| = |\Omega_0^{(i)}|$, where we have denoted the repeated set of frequencies common to all encoding gates as $\Omega_0^{(i)}$. Using the results on the maximum size of the spectrum of a κ-local Hamiltonian in Eq. (114), we can deduce that $T \leq 2^{\kappa-2}(2^\kappa - 1)$. We now quantify the number of different frequencies in $\Omega^{(i)}$ in terms of T. $N^{(i)}$ repetitions of the fixed Hamiltonian with frequencies $\Omega_0^{(i)}$ result in a set of frequencies that contains all possible combinations of $N^{(i)}$ vectors $\boldsymbol{\omega}_j$ from $\Omega_0^{(i)}$:

$$\Omega^{(i)} = \left\{ \sum_{j=1}^{N^{(i)}} \boldsymbol{\omega}_j \;\middle|\; \boldsymbol{\omega}_j \in \Omega_0^{(i)} \text{ for all } j \right\} \tag{117}$$

We can reformulate this by counting how often the $2T+1$ different elements of $\Omega_0^{(i)}$ are present in a particular instance of the above sum, and get

$$\Omega^{(i)} = \left\{ \sum_{\boldsymbol{\omega} \in \Omega_0^{(i)}} N_{\boldsymbol{\omega}} \boldsymbol{\omega} \;\middle|\; N_{\boldsymbol{\omega}} \geq 0,\ \sum_{\boldsymbol{\omega} \in \Omega_0^{(i)}} N_{\boldsymbol{\omega}} = N^{(i)} \right\}. \tag{118}$$

To bound the size of this set, we exploit the symmetry of the underlying frequencies $\delta_j \in \Delta(H^{(i)})$. Let us outline the idea: We will first count how we can distribute the number $N^{(i)}$ of repetitions over the different non-negative frequencies δ_j and then multiply this with the number of different frequencies that can be created by repeating δ_j and $-\delta_j$. To improve the scaling we get at the end, we will resort to a small trick and actually group the frequency 0, which we know to always be present in the spectrum, with the first other frequency, therefore considering the combinatoric problem of distributing $N^{(i)}$ "balls" over T distinguishable "bins" where some bins can be empty. The different possible ways to achieve this task are given by counting the *weak compositions* of $N^{(i)}$ into T parts, $\mathcal{C}(N^{(i)}, T)$. The number of such weak compositions is

$$|\mathcal{C}(N^{(i)}, T)| = \binom{N^{(i)} + T - 1}{N^{(i)}} \tag{119}$$

$$= \frac{(N^{(i)} + T - 1)!}{N^{(i)}!(T-1)!} \tag{120}$$

$$= \frac{(N^{(i)} + T - 1)(N^{(i)} + T - 2)\dots(N^{(i)} + 1)}{(T-1)!} \tag{121}$$

$$= \mathcal{O}((N^{(i)})^{T-1}). \tag{122}$$

We will denote such a composition as $(N_j^{(i)})_{j=1}^T \in \mathcal{C}(N^{(i)}, T)$. A simple counting argument reveals that there are $2N_1^{(i)} + 1$ possible sums with $N_1^{(i)}$ elements from the set $\{0, \delta_1, -\delta_1\}$ and $N_j^{(i)} + 1 \leq 2N_j^{(i)} + 1$ possible sums with $N_j^{(i)}$ elements from the set $\{\delta_j, -\delta_j\}$. We can therefore bound

$$|\Omega^{(i)}| \leq \sum_{(N_k^{(i)}) \in \mathcal{C}(N^{(i)}, T)} \prod_{k=1}^{T} \left(2N_k^{(i)} + 1\right) \tag{123}$$

$$\leq \sum_{(N_k^{(i)}) \in \mathcal{C}(N^{(i)}, T)} \left(\frac{2\sum_{k=1}^T N_k^{(i)}}{T} + 1\right)^T \tag{124}$$

$$\leq \sum_{(N_k^{(i)}) \in \mathcal{C}(N^{(i)}, T)} \left(\frac{2N^{(i)}}{T} + 1\right)^T \tag{125}$$

$$= |\mathcal{C}(N^{(i)}, T)| \left(\frac{2N^{(i)}}{T} + 1\right)^T \tag{126}$$

$$= \mathcal{O}((N^{(i)})^{2T-1}), \tag{127}$$

where we have used the arithmetic-geometric mean inequality to obtain the second inequality. From this inequality, we see that by repeating the same Hamiltonian for an encoding, we obtain a polynomial scaling in the number of repetitions whose exponent depends on the number of different frequencies generated by the repeated Hamiltonian.

Pauli encodings. Encodings performed with Hamiltonians that are a tensor product of Pauli operators, $H = \bigotimes_{k=1}^n P^{(k)}$ where $P^{(k)} \in \{\mathbb{I}, X, Y, Z\}$, have been analyzed in Ref. [33]. Therein, it was shown that $N^{(i)}$ repetitions of such encodings of arbitrary dimension will result in $|\Omega^{(i)}| = 2N^{(i)} + 1$.

Summary. We can easily connect the different upper bounds on $|\Omega^{(i)}|$ to upper bounds on $|\Omega|$ via the arithmetic-geometric mean inequality, i.e.,

$$\prod_{i=1}^{d} |\Omega^{(i)}| \leq \left(\sum_{i=1}^{d} \frac{|\Omega^{(i)}|}{d}\right)^d, \tag{128}$$

and by noting that, for $q \geq 1$,

$$\sum_{i=1}^{d} \frac{(N^{(i)})^q}{d} \leq \frac{\left(\sum_{i=1}^d N^{(i)}\right)^q}{d} = \frac{N^q}{d}. \tag{129}$$

Table 1: Scaling of the different upper bounds for the number of different frequencies for the encoding of a single parameter $|\Omega^{(i)}|$, as well as the associated bounds for the scaling of the number of different frequencies for the total data-encoding strategy, $|\Omega|$. $N^{(i)}$ denotes the number of gates used for encoding the input $x^{(i)}$, N denotes the total number of gates for all inputs.

Encoding strategy	Upper bound on $\lvert\Omega^{(i)}\rvert$	Upper bound on $\lvert\Omega\rvert$
Repetition of arbitrary Pauli encodings	$\mathcal{O}\left(N^{(i)}\right)$	$\mathcal{O}\left(\left(\frac{N}{d}\right)^d\right)$
Repetition of the same encoding with $2T+1$ frequencies	$\mathcal{O}\left((N^{(i)})^{2T-1}\right)$	$\mathcal{O}\left(\left(\frac{N^{2T-1}}{d}\right)^d\right)$
Repetition of the same κ-local encoding	$\mathcal{O}\left((N^{(i)})^{2^{\kappa+1}-1}\right)$	$\mathcal{O}\left(\left(\frac{N^{2^{\kappa+1}-1}}{d}\right)^d\right)$
Different κ-local encodings	$\mathcal{O}\left(2^{2\kappa N^{(i)}}\right)$	$\mathcal{O}\left(2^{2\kappa N}\right)$

Table 1 summarizes the different upper bounds on $|\Omega^{(i)}|$ for individual parameters $x^{(i)}$ derived in this section as well as the associated bounds on $|\Omega|$.

Given these results, we are finally in a position to provide a concrete answer to the first question posed in Section 2. More specifically, by substituting the upper bounds on $|\Omega|$ given in Table 1 into the generalization bounds for GTPs given in Section 5, we can obtain generalization bounds for PQC-based model classes which depend explicitly on architectural hyper-parameters associated with the data-encoding strategy. Recall that we denoted the function class associated with a particular set of parameters Θ, an encoding strategy $\mathcal{D}$ and an observable M, as $\mathcal{F}_{\Theta,\mathcal{D},M}$. We then obtain from Theorems 6 and 10 the following Corollary:

Corollary 14 (Generalization bound for PQCs—From Theorems 6 and 10)**.** *Let $d, m \in \mathbb{N}$. Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, c]$ be a bounded loss function such that $\mathbb{R} \ni z \mapsto \ell(y, z)$ is L-Lipschitz for all $y \in \mathbb{R}$. For any $\delta \in (0, 1)$ and for any probability measure P on $[0, 2\pi)^d \times \mathbb{R}$, with probability $\geq 1 - \delta$ over the choice of i.i.d. training data $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m \in ([0, 2\pi)^d \times \mathbb{R})^m$ of size m and every $f \in \mathcal{F}_{\Theta,\mathcal{D},M}$, where $\mathcal{D}$ is an encoding strategy with N gates in total, we have that,*

(a) if $\mathcal{D}$ denotes any data-encoding strategy consisting of Hamiltonians that are tensor products of Pauli operators,

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\frac{L\|M\|_\infty}{\sqrt{m}}\left(\frac{N}{d}\right)^{\frac{d}{2}} + c\sqrt{\frac{\log 1/\delta}{m}}\right), \tag{130}$$

(b) if $\mathcal{D}$ denotes any data-encoding strategy consisting of the same single Hamiltonian per data coordinate with T frequencies,

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\frac{L\|M\|_\infty}{\sqrt{m}}\left(\frac{N^{2T-1}}{d}\right)^{\frac{d}{2}} + c\sqrt{\frac{\log 1/\delta}{m}}\right), \tag{131}$$

(c) if $\mathcal{D}$ denotes any data-encoding strategy consisting of the same single κ-local Hamiltonian per data coordinate,

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\frac{L\|M\|_\infty}{\sqrt{m}}\left(\frac{N^{2^{\kappa+1}-1}}{d}\right)^{\frac{d}{2}} + c\sqrt{\frac{\log 1/\delta}{m}}\right), \tag{132}$$

(d) if $\mathcal{D}$ denotes any data-encoding strategy consisting of possibly different κ-local Hamiltonians per data coordinate,

$$R(f) - \hat{R}_S(f) \leq \tilde{\mathcal{O}}\left(\frac{L\|M\|_\infty}{\sqrt{m}}\, 2^{\kappa N} + c\sqrt{\frac{\log 1/\delta}{m}}\right). \tag{133}$$

While we consider only four specific data-encoding strategies in this corollary, the generalization bounds from Theorems 6 and 10 can in principle be applied to PQC-based models with *any* data-encoding strategy. To use the bounds, the corresponding $|\Omega(\mathcal{D})|$ has to be identified, which can then be readily combined with our generalization bounds for GTPs.

6.1 Comparison of data-encoding strategies from a generalization perspective

The results of the previous subsection give a concrete answer to the first question posed in Section 2, namely explicitly encoding-dependent generalization bounds for PQC-based models. However, recall from Section 2 that we also aimed to use such bounds to identify data-encoding strategies which give rise to a slow (polynomial) growth of model complexity with respect to increasingly complex data-encoding strategies, and therefore facilitate meaningful model selection via structural risk minimization. The results of the previous section now allow us to address this additional goal.

Given an assumption or constraint on the structure of the data-encoding Hamiltonians in a possible data-encoding strategy, the most natural data-encoding hyper-parameter for structural risk minimization is the number N of encoding Hamiltonians. We see that using either repeated Pauli Hamiltonians, a repeated (but fixed) κ-local Hamiltonian, or the repetition of a fixed Hamiltonian with $2T+1$ frequencies, leads to a complexity bound and generalization bound that scale polynomially with N. However, using N *different* κ-local data-encoding Hamiltonians can lead, in the worst case, to complexity upper bounds which scale exponentially with respect to N. In the latter case we stress, however, that these worst-case bounds are constructed using Hamiltonians designed to saturate the maximum possible number of frequency differences, and in many cases the complexity scaling with respect to N may be much slower. Additionally, while the polynomial generalization bounds we obtain for the first three data-encoding strategies give us hope in the possibility of meaningful structural risk minimization with respect to the number of data-encoding gates, our upper bounds on the generalization gap are not necessarily tight. Hence, we cannot rule out the possibility of better bounds for strategies consisting of many different Hamiltonians, which would facilitate the use of strucural risk minimization.

Additionally, while increasing the complexity of a data-encoding strategy by increasing N is a natural (and experimentally feasible) strategy, in principle one might also consider increasing either the locality κ or the number of frequencies T of the repeated data-encoding Hamiltonian. This would be particularly relevant in the realistic scenario where experimental constraints severely limit the number of data-encoding gates which can be used. However, apart from the potential experimental obstacles one would face in doing so, we note that while our complexity bounds are polynomial with respect to N (when keeping κ and T fixed), they are exponential (or doubly-exponential) with respect to κ and T respectively (when keeping N fixed). As such, given the generalization bounds we have obtained in this work, from the generalization and structural risk minimization perspective it makes the most sense to systematically increase the complexity of the data-encoding strategy by keeping κ and/or T constant, and increasing the number of data-encoding gates.

7 Discussion

As discussed in Section 2, the results from the previous section can be applied in a variety of ways. In particular, apart from the straightforward application of (probabilistically) bounding the generalization gap of an output hypothesis, or bounding the number of data samples required to guarantee an output hypothesis with a sufficiently small generalization gap, our results also facilitate the use of *structural risk minimization* with respect to architectural hyper-parameters related to the data-encoding strategy. We reiterate that the results obtained here should be viewed as *complementary* to many of the prior results discussed in Section 4. In particular, our results complement those which derive generalization bounds applicable to the same PQC-based hypothesis classes, but with explicit dependencies on architectural hyper-parameters which do not appear in our generalization bounds, such as depth, width, and total number of trainable gates.

More specifically, the generalization bounds of Section 6 allow one to use structural risk minimization to find the optimal setting for data-encoding hyper-parameters (in the sense of yielding an

output hypothesis with the smallest upper bound on true risk). However, they *do not* give any guidance as to how one should choose the remaining architectural hyper-parameters, and in particular those related to the trainable parts of the PQC. As such, a natural (and recommended) strategy is to use different available and applicable generalization bounds to perform "multi-dimensional structural risk minimization:" One can vary *all* architectural hyper-parameters for which one has a generalization bound, and evaluate each hyper-parameter setting with respect to an upper bound on the true risk obtained from a union bound over all existing applicable bounds. To make this more concrete, assume that we have a family of hypothesis classes $\{\mathcal{F}_{(k_1,k_2)}\}$, parametrized by two architectural hyper-parameters k_1 and k_2 (for example k_1 could be the number of encoding gates, and k_2 could be the number of trainable gates in a PQC based model). Additionally, let us assume that we have derived two different generalization bounds, one depending on k_1, the other depending on k_2. More concretely, assume that we have a function $g_1(k_1, m, \delta)$ and a function $g_2(k_2, m, \delta)$ such that, for all $i \in \{1, 2\}$, for all $\delta \in (0, 1)$, with probability $1 - \delta$ over $S \sim P^m$, for all $h \in \mathcal{F}_{(k_1,k_2)}$ we have that

$$R(h) \leq \hat{R}_S(h) + g_i(k_i, m, \delta). \tag{134}$$

Using a union bound, we can then straightforwardly combine these two results to obtain the following generalization bound: For all $\delta \in (0, 1)$, with probability $1 - \delta$ over $S \sim P^m$, for all $h \in \mathcal{F}_{(k_1,k_2)}$ we have that

$$R(h) \leq \hat{R}_S(h) + \min_i \left[g_i(k_i, m, \delta/2) \right]. \tag{135}$$

We see that we can perform structural risk minimization by varying both k_1 and k_2 and using $\min_i \left[g_i(k_i, m, \delta/2) \right]$ to calculate an upper bound on the true risk of the candidate hypothesis. The above argument can clearly be generalized to an arbitrary number of architectural hyper-parameters, and thereby yields a methodology for exploiting multiple existing generalization bounds for "multi-dimensional structural risk minimization."

While the approach we have just discussed certainly allows us to exploit existing complementary generalization bounds depending on different architectural hyperparameters, it is an interesting open question whether one can derive generalization bounds which depend *simultaneously* on multiple architectural hyper-parameters. In particular, it is of interest to understand whether one can in this way obtain generalization bounds, depending on multiple architectural hyper-parameters, which are *tighter* than the bounds obtained by taking a union bound over existing bounds, each of which depends only on a single hyper-parameter. A potential strategy for obtaining such bounds would be to better understand the effect of structural assumptions on the trainable part of a PQC architecture on the structure of the *coefficients* of the associated GTP representation. More concretely, while in this work we have focused on the frequency spectra of the GTPs, which are fully determined by the data-encoding strategy, the coefficients of the GTPs are determined by both the data-encoding strategy and the trainable part of the circuit. If one can characterize the implications of different circuit architectures on the structure of GTP coefficients, one could plausibly use refinements of the techniques presented in Section 5 to derive generalization bounds for the relevant GTPs that depend simultaneously on both the data-encoding strategy and complementary parameters of the circuit architecture. For example, certain PQC architectures may lead to GTP coefficients with a specific sparsity structure, or a constrained upper bound on a specific norm. Such a norm-specific bound may allow us to exploit the general p-norm extensions of our GTP bounds, mentioned in Remarks 8 and 12, to derive generalization bounds which also depend on the trainable circuit architecture.

Finally, we recall the potential shortcomings of *uniform* generalization bounds. In particular, in Ref. [36], the authors have shown both experimentally and analytically that sufficiently complex neural networks can achieve zero empirical risk for classification tasks with randomly assigned labels. As the true risk for such a learning problem can be no better than what would be achieved by random guessing, any *uniform* generalization bound for such a hypothesis class cannot offer any meaningful information in this complexity regime. More specifically, as uniform generalization bounds hold, by definition, for all hypotheses in the hypothesis class, and as there exist hypotheses which can achieve zero empirical risk even when generalization is not possible (i.e., when labels are selected randomly), such uniform bounds must be trivial.

It is, however, critical to emphasize that this finding applies only to *sufficiently complex* hypothesis classes. More specifically, they apply to models capable of achieving zero empirical risk even for completely unstructured data, which typically requires that the number of model parameters is at least as large as the number of elements in the training data set. As the number of parameters in a NISQ-regime PQC-based model is typically orders of magnitude less than the size of training data sets associated with "real-world" learning problems, it is unlikely that these known issues with uniform generalization bounds hinder the application of our uniform bounds to the analysis of currently available and near-term PQC-based hypothesis classes.

Despite this, it is important to keep these concerns in mind as the complexity of available PQC-based models increases. Consequently, there are a variety of natural open questions for future research: Firstly, can one replicate both the experimental and analytical aspects of Ref. [36] for PQC-based model classes? This would help to determine whether (or when) it is necessary to move beyond uniform generalization bounds for PQC-based models. In particular, from an experimental perspective, can one demonstrate the ability of a (sufficiently complex) PQC-based model class to achieve zero risk for a randomly-relabeled real-world classification task? Secondly, can one put an analytical bound on what is "sufficiently complex", i.e., how many model parameters are sufficient to ensure that for *any* training data set of size m, there *always* exists a hypothesis in the hypothesis class which can achieve zero empirical risk? Additionally, the shortcomings of uniform generalization bounds exposed in Ref. [36] have stimulated an explosion of research on non-uniform generalization bounds for highly complex neural network models [37]. It would be of interest to understand whether or how one can obtain non-uniform generalization bounds for PQC-based models, which would tighten the bounds obtained in this work in the future regime of high complexity.

8 Conclusion

In this work, we have derived Rademacher complexity and metric entropy bounds for PQC-based model classes. These depend explicitly on architectural hyper-parameters associated with the data-encoding strategy and are applicable to PQC-based models incorporating data re-uploading. By exploiting tools and techniques from statistical learning theory, we have then used these complexity bounds to obtain uniform generalization bounds, which allow to place a probabilistic upper-bound on the out-of-sample performance of any hypothesis, given its performance on the data. Additionally, we have used the obtained generalization bounds to compare data-encoding strategies from a generalization perspective and have discussed how, for certain data-encoding strategies, our generalization bounds may be used for model selection via structural risk minimization. We have stressed how the encoding-dependent generalization bounds obtained in this work should be viewed as *complementary* to existing complexity and generalization bounds for PQC-based models, which depend explicitly on architectural hyper-parameters to which our bounds are insensitive. More specifically, we have sketched in Section 7 how the combination of our bounds with existing works facilitates model selection via multi-dimensional structural risk minimization. Finally, as discussed in Section 7, it is important to acknowledge that the bounds we have obtained here are expected to be useful for PQC-based models in the "moderate-complexity" regime, i.e., for models parametrized by fewer parameters than the number of available data samples. However, in analogy with known results for classical model classes, these bounds may cease to be meaningful as the complexity of PQC-based models increases into an over-parametrized regime. Given this, we have also sketched in Section 7 a variety of open questions and directions for future research.

Acknowledgments

The authors would like to thank Alexander Nietner for insightful discussions and Maria Schuld and David Sutter for helpful feedback on an earlier draft. We would like to thank the Cluster of Excellence MATH+ (EF1-11), the BMWi (PlanQK), for which this work provides an understanding of models of quantum-enhanced machine learning, and the BMBF (Hybrid), for which this work helps with the design of quantum-classical hybrid models of quantum computing, for support. This work has also been supported by the DFG (CRC 183, project B01), the Einstein

Foundation (Einstein Research Unit on quantum devices) and by the EU's Horizon 2020 research and innovation programme under grant agreement No. 817482 (PASQuanS). M.C.C. gratefully acknowledges support from the TopMath Graduate Center of the TUM Graduate School at the Technical University of Munich, Germany, from the TopMath Program at the Elite Network of Bavaria, and from the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

A Auxiliary results from statistical learning theory

In this appendix, we collect some well known results from classical statistical learning theory that we make use of in our proofs.

Lemma 15 (Rademacher complexity progression (Theorem 2.15 in Ref. [35]))**.** *Let $a, b \in \mathbb{R}$ and $\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$ an L-Lipschitz function and assume $\mathcal{F}_0 \subseteq \mathbb{R}^{\mathcal{X}}$ is a set of functions that includes the 0 function. Also, let $\mathcal{F}$ be the following function class*

$$\mathcal{F} \coloneqq \left\{ \boldsymbol{x} \mapsto \tilde{\sigma}\left(v + \sum_{j=1}^{m} \omega_j f_j(\boldsymbol{x}) \right) \;\middle|\; |v| \leq a, \|\boldsymbol{\omega}\|_1 \leq b, \text{ and } f_j \in \mathcal{F}_0 \right\}. \tag{136}$$

Then, the empirical Rademacher complexity of $\mathcal{F}$ with respect to any point $\vec{\boldsymbol{x}} \in \mathcal{X}^m$ can be bounded in terms of the one of $\mathcal{F}_0$

$$\hat{\mathcal{R}}_{\vec{\boldsymbol{x}}}(\mathcal{F}) \leq L\left(\frac{a}{\sqrt{m}} + 2b\hat{\mathcal{R}}(\mathcal{F}_0) \right). \tag{137}$$

The 2 factor can be dropped if $\mathcal{F}_0 = -\mathcal{F}_0$.

Lemma 16 (Rademacher complexity of layered network (Corollary 2.11 in Ref. [35]))**.** *Let $a, b > 0$ and $\mathcal{X} \coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}^d \,\middle|\, \|\boldsymbol{x}\|_\infty \leq C \right\}$. Consider a neural network architecture with δ hidden layers that implements $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, and such that*

1. *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz and anti-symmetric.*
2. *For every neuron, the vector of weights $\boldsymbol{\omega}$ satisfies $\|\boldsymbol{\omega}\|_1 \leq b$.*
3. *For every neuron, the modulus of the bias is upper-bounded by a.*

Then, the empirical Rademacher complexity of $\mathcal{F}$ with respect to any point $\vec{\boldsymbol{x}} \in \mathcal{X}^m$ can be upper-bounded as

$$\hat{\mathcal{R}}_{\vec{\boldsymbol{x}}}(\mathcal{F}) \leq \frac{1}{\sqrt{m}} \left(C b^\delta \sqrt{2\log(2d)} + a \sum_{i=0}^{\delta-1} b^i \right). \tag{138}$$

Lemma 17 (Massart's Lemma [48])**.** *Let $N \in \mathbb{N}$. Let $A \subset \mathbb{R}^N$ be a finite set contained in a Euclidean ball of radius $r > 0$. Then*

$$\mathbb{E}_\sigma \left[\sup_{a \in A} \frac{1}{n} \sum_{i=1}^{N} \sigma_i a_i \right] \leq \frac{r\sqrt{2\log|A|}}{N}, \tag{139}$$

where the expectation is with respect to i.i.d. Rademacher random variables $\sigma_1, \ldots, \sigma_N$.

Lemma 18 (Talagrand's Lemma (going back to [46]; see also Lemma 5.7 in Ref. [20]))**.** *Let $\ell_1, \ldots, \ell_m : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz functions. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ be a class of real-valued functions on some data space $\mathcal{Z}$. Then, for any $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \in \mathcal{Z}$,*

$$\frac{1}{m}\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i \ell \circ f(\boldsymbol{z}_i) \right] \leq \frac{L}{m}\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}_\Omega^B} \sum_{i=1}^{m} \sigma_i f(\boldsymbol{z}_i) \right], \tag{140}$$

where the expectations are over i.i.d. Rademacher random variables $\sigma_1, \ldots, \sigma_m$.

Theorem 19 (Dudley's Theorem ([49]; see also Theorem 8.1.2 in Ref. [47] or Theorem 1.19 in Ref. [35])). *For a fixed vector $z \in \mathcal{Z}^m$ let $\mathcal{G}$ be a subset of the pseudo-metric space $(\mathbb{R}^{\mathcal{Z}}, \|\cdot\|_{2,z})$ and let $\gamma_0 \coloneqq \sup_{g\in\mathcal{G}} \|g\|_{2,z}$. Then the empirical Rademacher complexity $\hat{\mathcal{R}}_z(\mathcal{G})$ of $\mathcal{G}$ with respect to z can be upper-bounded as*

$$\hat{\mathcal{R}}_z(\mathcal{G}) \leq \inf_{\varepsilon\in[0,\frac{\gamma_0}{2})} \left\{ 4\varepsilon + \frac{12}{\sqrt{m}} \int_{\varepsilon}^{\gamma_0} \sqrt{\log \mathcal{N}(\mathcal{G}, \|\cdot\|_{2,z}, \beta)}\, \mathrm{d}\beta \right\}. \tag{141}$$

References

[1] V. Dunjko and H. J. Briegel, "Machine learning & artificial intelligence in the quantum domain: a review of recent progress", Rep. Prog. Phys. **81**, 074001 (2018) DOI: 10.1088/1361-6633/aab406.

[2] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning", Nature **549**, 195–202 (2017) DOI: 10.1038/nature23474.

[3] S. Arunachalam and R. de Wolf, "Guest column: a survey of quantum learning theory", SIGACT News **48**, DOI: 10.1145/3106700.3106710 (2017) DOI: 10.1145/3106700.3106710.

[4] N. Wiebe, A. Kapoor, and K. Svore, "Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning", Quant. Inf. Comp. **15**, 0318 (2015) DOI: 10.5555/2871393.2871400.

[5] S. Lloyd, M. Mohseni, and P. Rebentrost, "Quantum algorithms for supervised and unsupervised machine learning", arXiv:1307.0411 (2013).

[6] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, "On the quantum versus classical learnability of discrete distributions", Quantum **5**, 417 (2021) DOI: 10.22331/q-2021-03-23-417.

[7] Y. Liu, S. Arunachalam, and K. Temme, "A rigorous and robust quantum speed-up in supervised machine learning", Nature Physics **17**, 1013–1017 (2021) DOI: 10.1038/s41567-021-01287-z.

[8] F. Arute et al., "Quantum supremacy using a programmable superconducting processor", Nature **574**, 505–510 (2019) DOI: doi.org/10.1038/s41586-019-1666-5.

[9] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, "Noisy intermediate-scale quantum (NISQ) algorithms", arXiv:2101.08448 (2021).

[10] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, "The theory of variational hybrid quantum-classical algorithms", New J. Phys. **18**, 023023 (2016) DOI: 10.1088/1367-2630/18/2/023023.

[11] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "Variational quantum algorithms", Nature Reviews Physics **3**, 625–644 (2021) DOI: 10.1038/s42254-021-00348-9.

[12] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models", Quant. Sc. Tech. **4**, 043001 (2019) DOI: 10.1088/2058-9565/ab4eb5.

[13] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms", Adv. Quant. Tech. **2**, 1900070 (2019) DOI: 10.1002/qute.201900070.

[14] T. Hubregtsen, J. Pichlmeier, P. Stecher, and K. Bertels, "Evaluation of parameterized quantum circuits: on the relation between classification accuracy, expressibility, and entangling capability", Quant. Mach. Int. **3**, 1–19 (2021) DOI: 10.1007/s42484-021-00038-w.

[15] F. J. Gil Vidal and D. O. Theis, "Input redundancy for parameterized quantum circuits", Front. Phys. **8**, DOI: 10.3389/fphy.2020.00297 (2020) DOI: 10.3389/fphy.2020.00297.

[16] M. Schuld, "Quantum machine learning models are kernel methods", arXiv:2101.11020 (2021).

[17] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, "Data re-uploading for a universal quantum classifier", Quantum **4**, 226 (2020) DOI: 10.22331/q-2020-02-06-226.

[18] C. Bishop, *Pattern recognition and machine learning* (Springer, Berlin, 2006).

[19] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, Adaptive computation and machine learning (MIT Press, Cambridge, MA, 2002), 626 pp.

[20] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*, 2nd ed., Adaptive Computation and Machine Learning (MIT Press, Cambridge, MA, 2018), 504 pp.

[21] O. Bousquet and A. Elisseeff, "Stability and generalization", J. Mach. Learn. Res. **2**, 499–526 (2002) DOI: 10.1162/153244302760200704.

[22] N. Littlestone and M. Warmuth, "Relating data compression and learnability", Technical report, University of California Santa Cruz (1986).

[23] D. A. McAllester, "Some pac-bayesian theorems", Machine Learning **37**, 355–363 (1999) DOI: 10.1023/A:1007618624809.

[24] M. C. Caro and I. Datta, "Pseudo-dimension of quantum circuits", Quant. Mach. Int. **2**, 172 (2020) DOI: 10.1007/s42484-020-00027-5.

[25] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, "The power of quantum neural networks", Nature Computational Science **1**, 403–409 (2021) DOI: 10.1038/s43588-021-00084-1.

[26] K. Bu, D. E. Koh, L. Li, Q. Luo, and Y. Zhang, "On the statistical complexity of quantum circuits", arXiv:2101.06154 (2021).

[27] K. Bu, D. E. Koh, L. Li, Q. Luo, and Y. Zhang, "Effects of quantum resources on the statistical complexity of quantum circuits", arXiv:2102.03282 (2021).

[28] K. Bu, D. E. Koh, L. L., Q. Luo, and Y. Zhang, "Rademacher complexity of noisy quantum circuits", arXiv:2103.03139 (2021).

[29] Y. Du, Z. Tu, X. Yuan, and D. Tao, "An efficient measure for the expressivity of variational quantum algorithms", arXiv:2104.09961 (2021).

[30] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, "Power of data in quantum machine learning", Nature Comm. **12**, 1–9 (2021) DOI: 10.1038/s41467-021-22539-9.

[31] L. Banchi, J. Pereira, and S. Pirandola, "Generalization in quantum machine learning: A quantum information perspective", arXiv:2102.08991 (2021).

[32] C. Gyurik, D. van Vreumingen, and V. Dunjko, "Structural risk minimization for quantum linear classifiers", arXiv:2105.05566 (2021).

[33] M. Schuld, R. Sweke, and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models", Phys. Rev. A **103**, 032430 (2021) DOI: 10.1103/PhysRevA.103.032430.

[34] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: from theory to algorithms* (Cambridge University Press, 2014), DOI: 10.1017/CBO9781107298019.

[35] M. M Wolf, *Mathematical foundations of machine learning*, 2020.

[36] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization", arXiv:1611.03530 (2016).

[37] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them", arXiv:1912.02178 (2019).

[38] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities", Th. Prob. App. **16**, 264–280 (1971) DOI: 10.1137/1116025.

[39] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results", J. Mach. Learn. Res. **3**, 463–482 (2002) DOI: 10.5555/944919.944944.

[40] D. Pollard, *Convergence of stochastic processes*, Springer Series in Statistics (Springer, New York, NY, 1984), DOI: 10.1007/978-1-4612-5254-2.

[41] H. Abraham et al., *Qiskit: an open-source framework for quantum computing*, 2019, DOI: 10.5281/zenodo.2562110.

[42] Cirq Developers, *Cirq*, Mar. 5, 2021, DOI: 10.5281/zenodo.4586899.

[43] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, M. S. Alam, S. Ahmed, J. M. Arrazola, C. Blank, A. Delgado, S. Jahangiri, K. McKiernan, J. J. Meyer, Z. Niu, A. Száva, and N. Killoran, "PennyLane: automatic differentiation of hybrid quantum-classical computations", arXiv:1811.04968 (2020).

[44] C. M. Popescu, "Learning bounds for quantum circuits in the agnostic setting", Quantum Information Processing **20**, 1–24 (2021) DOI: 10.1007/s11128-021-03225-7.

[45] C.-C. Chen, M. Watabe, K. Shiba, M. Sogabe, K. Sakamoto, and T. Sogabe, "On the expressibility and overfitting of quantum circuit learning", ACM Transactions on Quantum Computing **2**, 1–24 (2021) DOI: 10.1145/3466797.

[46] M. Ledoux and M. Talagrand, *Probability in banach spaces: isoperimetry and processes* (Springer-Verlag, Berlin New York, 1991), DOI: 10.1007/978-3-642-20212-4.

[47] R. Vershynin, *High-dimensional probability: an introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, 2018), DOI: 10.1017/9781108231596.

[48] P. Massart, "Some applications of concentration inequalities to statistics", Annales de la Faculté des sciences de Toulouse : Mathématiques **Ser. 6, 9**, 245–303 (2000) DOI: 10.5802/afst.961.

[49] R. M. Dudley, *Uniform central limit theorems*, Cambridge Studies in Advanced Mathematics (Cambridge University Press, 1999), DOI: 10.1017/CBO9780511665622.

Appendix B

Further articles as principal author

B.1 Quantum learning Boolean linear functions w.r.t. product distributions

Quantum learning Boolean linear functions w.r.t. product distributions

Matthias C. Caro

Quantum Fourier sampling allows a quantum computer to sample from the probability distribution given by the squares of the Fourier coefficients of a Boolean function, assuming access to quantum superposition examples for the function. This tool serves as a subroutine in several quantum learning algorithms. However, the majority of Fourier-based quantum learning procedures are designed specifically for learning from uniform superpositions. In this work, we investigate the task of learning Boolean linear functions from quantum examples with non-uniform superposition weights, giving both sample complexity upper and lower bounds.

After the introductory Section 1, Section 2 contains the mathematical preliminaries for the remainder of the paper. In particular, in addition to recalling fundamental notions from quantum information and learning theory, respectively, there we also describe the basic concepts in classical biased Fourier analysis of Boolean functions. Moreover, we discuss the extension of quantum Fourier sampling to biased product distributions as well as the pretty good measurement as a useful tool for analysing success probabilities in distinguishing quantum states.

In Section 3, we describe the quantum learning problem. The goal is to exactly learn an unknown Boolean linear function $f^{(a)} : \{-1,1\}^n \to \{0,1\}$, which is computed by taking the inner product modulo 2 of an input n-bit string with an unknown n-bit string a. In our setting, a quantum learner has access to quantum examples of the form $\sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}\,|x, f^{(a)}(x)\rangle$, where $D_\mu(x) \coloneqq \prod_{i=1}^n (1+x_i\mu_i)/2$ defines a product distribution with bias vector $\mu \in [-1,1]^n$. As a useful subroutine in solving this problem, we propose a generalized Bernstein-Vazirani algorithm, Algorithm 2 in Section 4. By analyzing the biased Fourier coefficients of a Boolean linear function, we completely characterize the output distribution of biased quantum Fourier sampling when performed on a correspondingly biased superposition state (Theorem 2).

Section 5 contains our upper bounds on the quantum sample complexity of learning Boolean linear functions with respect to biased product distributions. On a high level, we obtain them by "amplifying" the success probability in the generalized Bernstein-Vazirani algorithm. Algorithm 3 in Subsection 5.1 describes our first amplification procedure, which is applicable for arbitrary (except full) bias. Theorem 3 shows that Algorithm 3 can exactly learn a Boolean linear function from $\mathcal{O}(\ln(n))$ biased superposition examples, with high success probability. In Subsection 5.2, we present an alternative amplified version of the generalized Bernstein-Vazirani algorithm as Algorithm 4. In the case of small bias, we prove in Theorem 4 that this procedure leads to an effectively n-independent sample complexity bound. In Appendix A, we prove analogues of Theorem 4 for the cases of noisy quantum training data and a noisy implementation of the involved quantum Fourier transforms. Moreover, Corollary 3 in Appendix A.3 gives a sample complexity upper bound for the case in which only a bound on the bias, but not the exact bias vector, is known in advance.

We complement these results with sample complexity lower bounds in Section 6. In Subsection 6.1, we demonstrate how to use an information-theoretic reasoning to recover the well-known sample complexity lower bound of $\Omega(n)$ for exactly learning an unknown Boolean linear function

from classical training data. Next, we turn to sample complexity lower bounds for the quantum case in Subsection 6.2. In the case of strong bias, we prove in Theorem 6 that $\Omega(\ln(n))$ quantum superposition examples are required to exactly learn an unknown Boolean linear function. For our proofs, we relate the learning problem to a quantum state discrimination task and establish lower bounds for the latter. We achieve this by bounding the success probability of the pretty good measurement, which requires us to perform a detailed analysis of the Gram matrix for the ensemble of possible quantum superposition examples.

I developed the idea for this project myself. I am solely responsible for the scientific content of this article. As the single author of this article, I am solely responsible for writing this article.

Note: This article is based on results from my Master's thesis.

Permission to include:

Matthias C. Caro.
Quantum learning Boolean linear functions w.r.t. product distributions.
Quantum Inf Process 19, 172 (2020). https://doi.org/10.1007/s11128-020-02661-1.

Springer www.springer.com

Permissions

Get permission to reuse Springer Nature content

Springer Nature is partnered with the Copyright Clearance Center to meet our customers' licensing and permissions needs.

Copyright Clearance Center's RightsLink® service makes it faster and easier to secure permission for the reuse of Springer Nature content to be published, for example, in a journal/magazine, book/textbook, coursepack, thesis/dissertation, annual report, newspaper, training materials, presentation/slide kit, promotional material, etc.

Simply visit SpringerLink and locate the desired content;

Go to the article or chapter page you wish to reuse content from. (Note: permissions are granted on the article or chapter level, not on the book or journal level). Scroll to the botton of the page, or locate via the side bar, the "Reprints and Permissions" link at the end of the chapter or article.

Select the way you would like to reuse the content;

Complete the form with details on your intended reuse. Please be as complete and specific as possible ao as not to delay your permission request;

Create an account if you haven't already. A RightsLink account is different than a SpringerLink account, and is necessary to receive a licence regardless of the permission fee. You will receive your licence via the email attached to your RightsLink receipt;

Accept the terms and conditions and you're done!

For questions about using the RightsLink service, please contact Customer Support at Copyright Clearance Center via phone +1-855-239-3415 or +1-978-646-2777 or email springernaturesupport@copyright.com.

How to obtain permission to reuse Springer Nature content not available online on SpringerLink

Requests for permission to reuse content (e.g. figure or table, abstract, text excerpts) from Springer Nature publications currently not available online must be submitted in writing. Please be as detailed and specific as possible about what, where, how much, and why you wish to reuse the content.

Your contacts to obtain permission for the reuse of material from:

- books: bookpermissions@springernature.com
- journals: journalpermissions@springernature.com

Author reuse

Please check the Copyright Transfer Statement (CTS) or Licence to Publish (LTP) that you have signed with Springer Nature to find further information about the reuse of your content.

Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.
Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]'

If you are any doubt about whether your intended re-use is covered, please contact journalpermissions@springernature.com for confirmation.

Self-Archiving

- Journal authors retain the right to self-archive the final accepted version of their manuscript. Please see our self-archiving policy for full details:
https://www.springer.com/gp/open-access/authors-rights/self-archiving-policy/2124

- Book authors please refer to the information on this link:
https://www.springer.com/gp/open-access/publication-policies/self-archiving-policy

Quantum Information Processing (2020) 19:172
https://doi.org/10.1007/s11128-020-02661-1

Quantum learning Boolean linear functions w.r.t. product distributions

Matthias C. Caro[1]

Received: 5 August 2019 / Accepted: 29 March 2020 / Published online: 20 April 2020

Abstract

The problem of learning Boolean linear functions from quantum examples w.r.t. the uniform distribution can be solved on a quantum computer using the Bernstein–Vazirani algorithm (Bernstein and Vazirani, in: Kosaraju (ed) Proceedings of the twenty-fifth annual ACM symposium on theory of computing, ACM, New York, 1993. https://doi.org/10.1145/167088.167097). A similar strategy can be applied in the case of noisy quantum training data, as was observed in Grilo et al. (Learning with errors is easy with quantum samples, 2017). However, extensions of these learning algorithms beyond the uniform distribution have not yet been studied. We employ the biased quantum Fourier transform introduced in Kanade et al. (Learning dnfs under product distributions via μ-biased quantum Fourier sampling, 2018) to develop efficient quantum algorithms for learning Boolean linear functions on n bits from quantum examples w.r.t. a biased product distribution. Our first procedure is applicable to any (except full) bias and requires $\mathcal{O}(\ln(n))$ quantum examples. The number of quantum examples used by our second algorithm is independent of n, but the strategy is applicable only for small bias. Moreover, we show that the second procedure is stable w.r.t. noisy training data and w.r.t. faulty quantum gates. This also enables us to solve a version of the learning problem in which the underlying distribution is not known in advance. Finally, we prove lower bounds on the classical and quantum sample complexities of the learning problem. Whereas classically, $\Omega(n)$ examples are necessary independently of the bias, we are able to establish a quantum sample complexity lower bound of $\Omega(\ln(n))$ only under an assumption of large bias. Nevertheless, this allows for a discussion of the performance of our suggested learning algorithms w.r.t. sample complexity. With our analysis, we contribute to a more quantitative understanding of the power and limitations of quantum training data for learning classical functions.

Keywords Computational learning theory · Exact learning · Quantum Fourier learning

Matthias C. Caro
caro@ma.tum.de

Extended author information available on the last page of the article

1 Introduction

The origins of the fields of machine learning as well as quantum information and computation both lie in the 1980s. The arguably most influential learning model, namely the PAC ("probably approximately correct") model, was introduced by Valiant in 1984 [26] with which the problem of learning was given a rigorous mathematical framework. Around the same time, Benioff [7] and Feynman presented the idea of quantum computers [12] to the public and thus gave the starting signal for important innovations at the intersection of computer science, information theory and quantum theory. Both learning theory and quantum computation promise new realms of computation in which tasks that seem insurmountable from the perspective of classical computation become feasible. The first has already proved its practical worth and is indispensable for modern-world big data applications, the latter is not yet as practically relevant but much work is invested to make the promises of quantum computation a reality. The interested reader is referred to [20,25] for an introduction to statistical learning and quantum computation and information, respectively.

Considering the increasing importance of machine learning and quantum computation, attempting a merger of the two seems a natural step to take and the first step in this direction was taken already in [10]. The field of quantum learning has received growing attention over the last few years and by now some settings are known in which quantum training data and the ability to perform quantum computation can be advantageous for learning problems from an information-theoretic as well as from a computational perspective, in particular for learning problems with fixed underlying distribution (see, e.g., [3] for an overview). It was, however, shown in [4] that no such information-theoretic advantage can be obtained in the (distribution-independent) quantum PAC model (based on [10]) compared to the classical PAC model (introduced in [26]).

One of the early examples of the aptness of quantum computation for learning problems is the task of learning Boolean linear functions w.r.t. the uniform distribution via the Bernstein–Vazirani algorithm presented in [8]. Whereas this task of identifying an unknown n-bit string classically requires a number of examples growing (at least) linearly with n, a bound on the sufficient number of copies of the quantum example state independent of n can be established. This approach was taken up in [13] where it is shown that, essentially, the Bernstein–Vazirani-based learning method is also viable if the training data is noisy. However, also this analysis is restricted to quantum training data arising from the uniform distribution. The same limiting assumption was also made in [10] for learning Disjunctive Normal Forms and in this context an extension to product distributions was achieved in [17].

Hence, a next direction to go is building up on the reasoning of [17] to extend the applicability of quantum learning procedures for linear functions to more general distributions. The analysis hereby differs from the one for DNFs because no concentration results for the biased Fourier spectrum of a linear function are available. Moreover, whereas many studies of specific quantum learning tasks focus on providing explicit learning procedures yielding a better performance than known classical algorithms, we complement our learning algorithms with lower bounds on the size of the training

data for a comparison to the best classical procedure and for a discussion of optimality among possible quantum strategies.

1.1 Overview over the results

The task of learning linear functions has already served as a toy model for quantum speed-ups in the early days of quantum computing. We describe possible generalizations of known results in different scenarios. First, in Theorem 3 we exhibit a Fourier-sampling-based algorithm which learns Boolean linear functions on n inputs from $\mathcal{O}(\ln(n))$ quantum examples arising from a c-bounded product distribution D_μ. (Classically, it is known that $\Omega(n)$ examples are required.) Moreover, for a bias vector μ satisfying $|\mu_i| \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ for all i, this can be reduced to $\mathcal{O}(1)$ quantum examples (Theorem 4). We also show that this reduction to a constant number of quantum examples is not possible for arbitrary product distributions by giving quantum sample complexity lower bounds in Theorem 6.

In Theorem 8, we exhibit a noise bound for quantum examples arising from a product distribution D_μ with $|\mu_i| \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ for all i but corrupted by noise which guarantees that $\mathcal{O}(1)$ quantum examples still suffice for learning. Under milder assumptions on the noise, a $\mathcal{O}(\ln(n))$ upper bound on the sample complexity is given. Similarly, faulty quantum gates can be tolerated in our learning algorithm. Based on this observation, we construct a quantum learning algorithm without prior knowledge of the underlying distribution which requires $\mathcal{O}(n^2)$ quantum examples by first estimating the bias vector classically (Corollary 3).

1.2 Related work

The (classical) problem of learning linear functions from randomly drawn examples in the presence of noise was studied in [9] (over the field $\mathbb{F}_2$) as well as in [22] (over a field $\mathbb{F}_q$ for q prime). The latter of these two works also established the relevance of this learning problem for cryptography by connecting it to certain lattice problems. A different model for learning linear functions is studied in [16], where the training data is not assumed to be noisy but instead only partial information about the function values is revealed.

The quantum PAC model was introduced in [10], where it was employed for learning DNF formulae w.r.t. the uniform distribution using a quantum example oracle. This was extended to product distributions by [17]. On the basis of this notion of quantum examples, the known Bernstein–Vazirani algorithm [8] can be reinterpreted as giving rise to a quantum learning algorithm for linear functions. This interpretation is explicitly given and further elaborated upon for the case of noisy training data in [11] (for $q = 2$) and in [13] (for general primes q). Cross et al. [11] established that, whereas the learning parity problem without noise is feasible both for classical and quantum computation, the learning parity with noise problem is widely believed to be classically intractable but remains feasible for quantum computers, where the runtime depends only logarithmically on the number of qubits. This quantum advantage for

noisy systems was demonstrated experimentally in [23]. Grilo et al. [13] extends this analysis to general fields and a broader class of noise models and obtains that also for that scenario, learning linear functions from noisy data is feasible for quantum computers; however, their runtime bound is polynomial in the number of subsystems. In [5], the class of juntas is found to also allow for efficient quantum learning. The framework of Fourier-based quantum exact learning is shown to be efficiently applicable more generally also to Fourier-sparse functions in [1]. Limitations of the power of quantum computation for learning have been studied in a series of papers culminating in [4] and more recently also in [2]. The former work shows that without prior restrictions on the underlying probability distribution, quantum examples are not more powerful than classical examples. The latter work demonstrates that, assuming quantum hardness of the learning with errors problem from classical examples, the class of shallow circuits is hard to learn from quantum examples.

Aside from the task of learning from examples, also the problem of learning from membership queries, both classical and quantum, is well studied. For instance, [24] established a polynomial relation between the number of required quantum versus required classical queries, which was recently improved upon in [1]. Also, [19] uses quantum membership queries for learning multilinear polynomials more efficiently than is classically possible.

1.3 Structure of the paper

The paper is structured in the following way. In Sect. 2, we introduce the well-known notions from classical learning, quantum computation and Boolean Fourier analysis required for our purposes as well as the prototypic learning algorithm which motivates our procedures. Section 3 consists of a description of the learning task to be considered. This is followed by a generalization of the Bernstein–Vazirani algorithm to product distributions in Sect. 4. In the next section, this is used to develop two quantum algorithms for solving our problem. ("Appendix A" contains a stability analysis of the second of the two procedures w.r.t. noise in training data and computation.) In Sect. 6, we establish sample complexity lower bounds complementing the upper bounds implied by the algorithms of Sect. 5. Finally, we conclude with some open questions and the references.

2 Preliminaries

2.1 Basics of quantum information and computation

We first define some of the fundamental objects of quantum information theory, albeit restricted to those required in our discussion. For the purpose of our presentation, we will consider a pure n-qubit quantum state to be represented by a state vector $|\psi\rangle \in \mathbb{C}^{2^n}$ (in Dirac notation). Such a state encodes measurement probabilities in the following way: If $\{|b_i\rangle\}_{i=1}^{2^n}$ is an orthonormal basis of $\mathbb{C}^{2^n}$, then there corresponds a measurement to this basis and the probability of observing outcome i for a system in

state $|\psi\rangle$ is given by $|\langle b_i|\psi\rangle|^2$. Finally, when considering multiple subsystems we will denote the composite state by the tensor product, i.e., if the first system is in state $|\psi\rangle$ and the second in state $|\phi\rangle$, the composite system is in state $|\psi, \varphi\rangle := |\psi\rangle \otimes |\phi\rangle$.

Quantum computation now consists in evolution of quantum states. Performing a computational step on an n-qubit state corresponds to applying an $2^n \times 2^n$ unitary transformation to the current quantum state. (The most relevant example of such unitary gates in our context will be the (biased) quantum Fourier transform discussed in more detail in Sect. 2.4.) As the outcome of a quantum computation is supposed to be classical, as final step of our computation we perform a measurement such that the final output will be a sample from the corresponding measurement statistics.

We will also use some standard notions from (quantum) information theory. For example, we denote the Shannon entropy of a random variable X by $H(X)$, the conditional entropy of a random variable X given Y as $H(X|Y)$ and the mutual information between random variables X and Y as $I(X : Y)$. Similarly, the von Neumann entropy of a quantum state ρ will be denoted as $S(\rho)$ and the mutual information for a bipartite quantum state ρ_{AB} as $I(\rho_{AB}) = I(A : B)$. Standard results on these quantities which will enter our discussion can, e.g., be found in [20].

2.2 Basics of learning theory

Next we describe the model of exact learning. In classical exact learning for an input space $\mathcal{X}$, a target space $\{0, 1\}$, and a concept class $\mathcal{F} \subset \{0, 1\}^{\mathcal{X}}$, a learning algorithm receives as input labeled training data $\{(x_i, f(x_i))\}_{i=1}^m$ for some (to the learner) unknown $f \in \mathcal{F}$, where the x_i are drawn independently according to some probability distribution D on $\mathcal{X}$ which is known to the learner. The goal of the learner is to exactly reproduce the unknown function f from such training examples with high success probability.

We can formalize this as follows: We call a concept class $\mathcal{F}$ exactly learnable if there exists a learning algorithm $\mathcal{A}$ and a map $m_{\mathcal{F}} : (0, 1) \to \mathbb{N}$ s.t. for every $D \in \mathrm{Prob}(X)$ (where $\mathrm{Prob}(X)$ is the set of all probability measures on X), $f \in \mathcal{F}$ and $\delta \in (0, 1)$, running $\mathcal{A}$ on training data of size $m \geq m_{\mathcal{F}}(\delta)$ drawn according to D and f with probability $\geq 1 - \delta$ (w.r.t. the choice of training data) yields a hypothesis h s.t. $h(x) = f(x)$ for all $x \in \mathcal{X}$. The smallest such map $m_{\mathcal{F}}$ is called sample complexity of exactly learning $\mathcal{F}$.

Note that this definition of learning captures the information-theoretic challenge of the learning problem in the sample complexity, but it does not refer to the computational complexity of learning. The focus on sample complexity is typical in statistical learning theory. Hence, also our results will be formulated in terms of sample complexity bounds. As we give explicit algorithms, these results directly imply bounds on the computational complexity; however, we will not discuss them in any detail.

Note also that the exact learning model differs from the well-known PAC ("probably approximately correct"), introduced by [26], in two ways. First, whereas the PAC model only requires to approximate the unknown function with high probability, we require to reproduce it exactly; in other words, we set the accuracy in PAC learning

to 0. Second, whereas in the PAC scenario the learner does not know the underlying distribution, we assume it to be fixed and known in advance. A short discussion on how to relax this restriction can be found in Sect. A.3.

The quantum exact learning model differs from the classical model in the form of the training data and the allowed form of computation. Namely, in quantum exact learning, the training data consists of m copies of the quantum example state $|\psi_f\rangle = \sum_{x\in\mathcal{X}} \sqrt{D(x)}|x,\ f(x)\rangle$, and this training data is processed by quantum computational steps. With this small change, the above definition of exact learnability and sample complexity now carry over analogously.

We conclude this introduction with a concentration result that has proven to be useful throughout learning theory.

Lemma 1 (Hoeffding's Inequality [15], compare also Theorem 2.2.6 in [27])
Let $Z_1, \ldots, Z_n$ be real-valued independent random variables taking values in closed and bounded intervals $[a_i, b_i]$, respectively. Then for every $\varepsilon > 0$

$$\mathbb{P}\left[\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i] \geq \varepsilon\right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right).$$

This directly implies (after replacing Z_i with $-Z_i$) that

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right| \geq \varepsilon\right] \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right).$$

2.3 μ-biased Fourier analysis of Boolean functions

We now give the basic ingredients of μ-biased Fourier analysis over the Boolean cube $\{-1, 1\}^n$. For more details, the reader is referred to [21].
For a bias vector $\mu \in [-1, 1]^n$, define the μ-biased product distribution D_μ on $\{-1, 1\}^n$ via

$$D_\mu(x) := \left(\prod_{i:x_i=1} \frac{1+\mu_i}{2}\right)\left(\prod_{i:x_i=-1} \frac{1-\mu_i}{2}\right) = \prod_{1\leq i\leq n} \frac{1+x_i\mu_i}{2}, \quad x \in \{-1, 1\}^n.$$

Thus, a positive μ_i tells us that at the ith position the distribution is biased towards $+1$, a negative μ_i tells us that at the ith position the distribution is biased towards -1. For $\mu = 0 \ldots 0$, we simply obtain the uniform distribution on $\{-1, 1\}^n$. The absolute value of μ_i quantifies the strength of the bias in the ith component. We call D_μ c-bounded, for $c \in (0, 1]$, if $\mu \in [-1 + c, 1 - c]^n$. Assuming the underlying product distribution to be c-bounded thus corresponds to assuming that the bias is not arbitrarily strong. Hence, we will in the following express notions of "small" or "large" bias either in terms of the bias vector μ or in terms of the c-boundedness constant.

For Fourier analysis, we now need an orthonormal basis for the function space $\mathbb{R}^{\{-1,1\}^n}$ w.r.t. the inner product $\langle ., . \rangle_\mu$ defined by

$$\langle f, g \rangle_\mu = \mathbb{E}_{D_\mu}[fg] = \sum_{x \in \{-1,1\}^n} f(x)g(x)D_\mu(x).$$

One can show (using the product structure to reduce to the case $n = 1$) that such an orthonormal basis is given by $\{\phi_{\mu,j}\}_{j \in \{0,1\}^n}$ with $\phi_{\mu,j}(x) = \prod_{i:j_i=1} \frac{x_i - \mu_i}{\sqrt{1-\mu_i^2}}$. For a function $f : \{-1, 1\}^n \to \{-1, 1\}$ this now gives a representation $f(x) = \sum_{j \in \{0,1\}^n} \hat{f}_\mu(j)\phi_{\mu,j}(x)$ with $\hat{f}_\mu(j) := \langle f, \phi_{\mu,j} \rangle_\mu$. For $\mu = 0 \ldots 0$, we recover the well-known orthonormal basis consisting of $\chi_j(x) = (-1)^{j \cdot x}$ from standard Fourier analysis over the Boolean cube.

2.4 μ-biased quantum Fourier sampling

We now turn to the description of the quantum algorithm for μ-biased quantum Fourier sampling which constitutes the basic ingredient of our learning algorithms and which, to our knowledge, was first presented in [17]. There the authors demonstrate that the μ-biased Fourier transform for a c-bounded D_μ with $c \in (0, 1]$ can be implemented on a quantum computer as the n-qubit μ-biased quantum Fourier transform: For $x \in \{-1, 1\}^n$,

$$H_\mu^n |x\rangle = H_\mu \otimes \ldots \otimes H_\mu |x_1, \ldots, x_n\rangle = \sum_{j \in \{0,1\}^n} \sqrt{D_\mu(x)}\phi_{\mu,j}(x)|j\rangle.$$

In the same way as the unbiased quantum Fourier transform can be used for quantum Fourier sampling, this μ-biased version now yields a procedure to sample from the μ-biased Fourier spectrum of a function using a quantum computer. We describe the corresponding procedure in Algorithm 1.

Algorithm 1 μ-biased Quantum Fourier Sampling

Input: $|\psi_f\rangle = \sum_{x \in \{-1,1\}^n} \sqrt{D_\mu(x)}|x, f(x)\rangle$ for a function $f : \{-1, 1\}^n \to \{0, 1\}$

Output: $j \in \{0, 1\}^n$ with probability $\left(\hat{g}_\mu(j)\right)^2$, where the function $g : \{-1, 1\}^n \to \{-1, 1\}$ is defined as $g(x) = (-1)^{f(x)}$.

Success Probability: $\frac{1}{2}$

1: Perform the μ-biased QFT H_μ on the first n qubits, obtain the state $(H_\mu \otimes \mathbb{1})|\psi_f\rangle$.
2: Perform a Hadamard gate on the last qubit, obtain the state $(H_\mu \otimes H)|\psi_f\rangle$.
3: Measure each qubit in the computational basis and observe outcome $j = j_1 \ldots j_{n+1}$.
4: **if** $j_{n+1} = 0$ **then** ▷ This corresponds to a failure of the sampling algorithm.
5: Output $o \leftarrow \perp$ and end computation.
6: **else if** $j_{n+1} = 1$ **then** ▷ This corresponds to a success of the sampling algorithm.
7: Output $o \leftarrow j_1 \ldots j_n$ and end computation.
8: **end if**

One can show that this algorithm indeed works as claimed by analyzing the transformation of the quantum state throughout the steps algorithm and making use of the orthonormality of the basis. This is the content of the following

Lemma 2 (Lemma 3 in [17])

Denote $g : \{-1, 1\}^n \to \{-1, 1\}$, $g(x) = (-1)^{f(x)}$. *Then with probability* $\frac{(\hat{g}_\mu(j))^2}{2}$, *Algorithm 1 outputs the string* $j \in \{0, 1\}^n$.

Proof The proof can be found in [17], we reproducce it in "Appendix B." □

This result allows us to generalize results based on quantum Fourier sampling w.r.t. the uniform distribution. In particular, we will apply it to obtain a generalization of the Bernstein–Vazirani algorithm.

2.5 The pretty good measurement

A basic problem in quantum information is that of distinguishing quantum states. We now describe a useful tool in this context, namely a measurement that is guaranteed to have a "pretty good" success probability to correctly identify an unknown state from a known ensemble.

Suppose that Alice (A) chooses one among m pure states $|\psi_i\rangle \in \mathbb{C}^d$ according to probabilities $p_i \in [0, 1]$, where $p_i \geq 0$ and $\sum_{i=1}^m p_i = 1$ and then sends the state to Bob (B). B wants to identify the state by performing a POVM measurement $\mathcal{A}$. Let $\mathcal{E} = \{(p_i, |\psi_i\rangle)\}_{i=1\ldots,m}$ be the ensemble describing A's preparation procedure, denote B's optimal success probability by $P^{opt} := \max_{POVM\ \mathcal{A}} P^{\mathcal{A}}$, where $P^{\mathcal{A}} := \sum_{i=1}^m p_i \langle\psi_i|A_i|\psi_i\rangle$ for a POVM $\mathcal{A} = \{A_i\}_{i=1,\ldots,m}$. Hausladen and Wootters [14] suggested a canonical form for a measurement for state discrimination, which is now usually referred to as the "pretty good measurement" (PGM) corresponding to the ensemble $\mathcal{E}$. It is defined in the following way:

First let $|\psi_i'\rangle := \sqrt{p_i}|\psi_i\rangle$ be the states renormalized according to their respective probabilities. The density operator of the ensemble $\mathcal{E}$ is $\rho := \sum_{i=1}^m p_i|\psi_i\rangle\langle\psi_i| = \sum_{i=1}^m |\psi_i'\rangle\langle\psi_i'|$. Now define $|\varphi_i\rangle := \rho^{-\frac{1}{2}}|\psi_i\rangle$, where the inverse square root is taken only over nonzero eigenvalues of ρ. Now the PGM is $\mathcal{A}^{PGM} = \{|\varphi_i\rangle\langle\varphi_i|\}_{i=1,\ldots,m}$. (Observe that this is indeed a valid POVM, even a projection-valued measure (PVM), because $\sum_{i=1}^m |\varphi_i\rangle\langle\varphi_i| = \rho^{-\frac{1}{2}}\rho\rho^{-\frac{1}{2}} = \mathbb{1}_d$.)

The "pretty good" performance of the PGM was proved in [6]:

Theorem 1 *For the PGM measurement defined above it holds that*

$$P^{opt}(\mathcal{E})^2 \leq P^{PGM}(\mathcal{E}) \leq P^{opt}(\mathcal{E}).$$

Another useful property of the PGM is that the corresponding success probability can be computed from the Gram matrix of the ensemble as follows:

Lemma 3 *The success probability for the PGM measurement for an ensemble* $\mathcal{E} = \{(p_i, |\psi_i\rangle)\}_{i=1\ldots,m}$ *can be written as*

$$P^{PGM}(\mathcal{E}) = \sum_{i=1}^{m} \sqrt{G}(i,i)^2,$$

where G *is the Gram matrix with entries* $G(i,j) = \sqrt{p_i p_j}\langle\psi_i|\psi_j\rangle$ *for* $1 \leq i, j \leq m$.

Proof This result can be shown by direct computation using the definition of the PGM and the uniqueness of the positive square root of a positive matrix. □

3 The learning problem

We now describe the learning task which we aim to understand. For $a \in \{0,1\}^n$, define

$$f^{(a)} : \{-1,1\}^n \to \{0,1\}, \quad f^{(a)}(x) := \sum_{i=1}^{n} a_i \frac{1-x_i}{2} \pmod 2.$$

When we observe that $\frac{1-x_i}{2}$ is simply the bit-description of x_i, it becomes clear that $f^{(a)}$ computes the parity of the entries of the bit-description of x_i at the positions at which a has a 1-entry. To ease readability, we will write $\tilde{x}_i = \frac{1-x_i}{2}$.

The classical task which inspires our problem is the following: Given a set of m labeled examples $S = \{(x_i, f^{(a)}(x_i))\}_{i=1}^m$, where the x_i are drawn i.i.d. according to D_μ, determine the string a with high success probability. Here, we assume prior knowledge of the underlying distribution and that the underlying distribution is a c-bounded product distribution as introduced in Sect. 2.4. This means that we are considering a problem of exact learning from examples with instances drawn from a distribution that is known to the learner in advance.

Classically, as we show in Sect. 6, successfully solving the task requires a number of examples that grows at least linearly in n. If we consider a version of this problem with noisy training data, then known classical algorithms perform worse both w.r.t. sample complexity and running time. For example, [18] exhibits an algorithm with polynomial (superlinear) sample complexity but barely subexponential runtime (both w.r.t. n).

The step to the quantum version of this problem now is the same as from classical to quantum exact learning. This means that training data is given as m copies of the quantum example state $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$ and the learner is allowed to use quantum computation to process the training data. The goal of the quantum learner remains that of outputting the unknown string a with high success probability.

4 A generalized Bernstein–Vazirani algorithm

To understand how μ-biased quantum Fourier sampling can help us with this learning problem, we first compute the μ-biased Fourier coefficients of $g^{(a)} := (-1)^{f^{(a)}}$, with $f^{(a)}$ for $a \in \{0,1\}^n$ the linear functions defined in Sect. 3.

Lemma 4 *Let $a \in \{0,1\}^n$, $g^{(a)} := (-1)^{f^{(a)}}$ and $\mu \in (-1,1)^n$. Then the μ-biased Fourier coefficients of $g^{(a)}$ satisfy:*

(i) *If $\exists\, 1 \le i \le n$ s.t. $a_i = 0 \neq j_i$, then $\hat{g}_\mu^{(a)}(j) = 0$.*
(ii) *If for all $1 \le i \le n$ s.t. $a_i = 0$ also $j_i = 0$, then*

$$\hat{g}_\mu^{(a)}(j) = \left(\prod_{l:a_l=1\neq j_l} \mu_l \right) \left(\prod_{l:a_l=1=j_l} \sqrt{1-\mu_l^2} \right).$$

We can reformulate this as

$$\hat{g}_\mu^{(a)}(j) = \left(\prod_{l:a_l=0} (1-j_l) \right) \left(\prod_{l:a_l=1} \left((1-j_l)\mu_l + j_l\sqrt{1-\mu_l^2} \right) \right), \quad j \in \{0,1\}^n.$$

Proof We first observe that all the "objects of interest," namely the probability distribution D_μ, the basis functions $\phi_{\mu,j}$, and the target function $\hat{g}_\mu^{(a)}$, factorize. This now implies that also the μ-biased Fourier coefficients factorize, i.e., we have

$$\hat{g}_\mu^{(a_1\ldots a_n)}(j_1\ldots j_n) = \prod_{i=1}^{n} \mathbb{E}_{D_{\mu_i}}[\phi_{\mu_i,j_i}(x_i)\cdot(-1)^{a_i\cdot\tilde{x}_i}].$$

Therefore we only have to study the case $n = 1$ in detail and the general result then follows. In this case, we have $f^{(a)}(x) = a\tilde{x}$, $g^{(a)}(x) = (-1)^{a\tilde{x}}$ for $\tilde{x} = \frac{1-x}{2}$, $\phi_{\mu,0}(x) = 1$, and $\phi_{\mu,1}(x) = \frac{x-\mu}{\sqrt{1-\mu^2}}$. (We leave out unnecessary indices to improve readability.) We compute

$$\hat{g}_\mu^{(a)}(j) = \mathbb{E}_{D_\mu}[(-1)^{a\tilde{x}}\phi_{\mu,j}(x)] = \frac{1+\mu}{2}\cdot 1\cdot\phi_{\mu,j}(1) + \frac{1-\mu}{2}\cdot(-1)^a\cdot\phi_{\mu,j}(-1).$$

By plugging in we now obtain

$$\hat{g}_\mu^{(0)}(0) = 1, \quad \hat{g}_\mu^{(0)}(1) = 0, \quad \hat{g}_\mu^{(1)}(0) = \mu, \quad \hat{g}_\mu^{(1)}(1) = \sqrt{1-\mu^2},$$

which is exactly the claim for $n = 1$. □

For clarity, we write down explicitly the algorithm which we obtain as a generalization of the Bernstein–Vazirani algorithm to a μ-biased product distribution as

Algorithm 2. The generalization compared to the standard Bernstein–Vazirani algorithm consists only in going from the uniform to a more general product distribution, which gives rise to different observation probabilities.

Algorithm 2 Generalized Bernstein–Vazirani algorithm

Input: $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$ for $a \in \{0,1\}^n$, and $\mu \in [-1,1]^n$

Output: $o \in \{0,1\}^n$ with probability

$$\left(\prod_{l:a_l=0}(1-o_l)\right)\left(\prod_{l:a_l=1}\left((1-o_l)\mu_l^2 + o_l(1-\mu_l^2)\right)\right)$$

Success Probability: $\frac{1}{2}$

1: Perform the μ-biased QFT H_μ on the first n qubits, obtain the state $(H_\mu \otimes \mathbb{1})|\psi_a\rangle$.
2: Perform a Hadamard gate on the last qubit, obtain the state $(H_\mu \otimes H)|\psi_a\rangle$.
3: Measure each qubit in the computational basis and observe outcome $j = j_1 \dots j_{n+1}$.
4: **if** $j_{n+1} = 0$ **then** ▷ This corresponds to a failure of the algorithm.
5: Output $o = \perp$.
6: **else if** $j_{n+1} = 1$ **then** ▷ This corresponds to a success of the algorithm.
7: Output $o = j_1 \dots j_n$.
8: **end if**

We now show that the output probabilities of Algorithm 2 are as claimed in its description. This follows directly by combining Lemma 2 on the workings of μ-biased quantum Fourier sampling with Lemma 4 on the μ-biased Fourier coefficients of our target functions and is the content of the following

Theorem 2 *Let* $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$ *be a quantum example state, with* $a \in \{0,1\}^n$ *and* $\mu \in (-1,1)^n$. *Then step* 3 *of Algorithm* 2 *provides an outcome* $|j_1 \dots j_{n+1}\rangle$ *with the following properties:*

(i) $\mathbb{P}[j_{n+1} = 0] = \frac{1}{2} = \mathbb{P}[j_{n+1} = 1]$,

(ii) $\mathbb{P}[j_1 \dots j_n = a | j_{n+1} = 1] = \prod_{l:a_l=1}(1-\mu_l^2)$,

(iii) *for* $o \neq a$:

$$\mathbb{P}[j_1 \dots j_n = o | j_{n+1} = 1] = \prod_{l:a_l=0}(1-o_l) \cdot \prod_{l:a_l=1}\left((1-o_l)\mu_l^2 + o_l(1-\mu_l^2)\right),$$

(iv) $\mathbb{P}[\exists 1 \le i \le n : a_i = 0 \neq j_i | j_{n+1} = 1] = 0$, *and*

(v) $\mathbb{P}[\exists 1 \le i \le n : a_i = 1 \neq j_i | j_{n+1} = 1] \le \sum_{i=1}^n \mu_i^2$. *In particular, if* D_μ *is* c*-bounded, then* $\mathbb{P}[\exists 1 \le i \le n : a_i = 1 \neq j_i | j_{n+1} = 1] \le n(1-c)^2$.

Note that (v) can be trivial if the bias is too strong. This observation already hints at why we later use different procedures for arbitrary and for small bias.

We also want to point out that in the case of no bias (i.e., $\mu = 0$), Algorithm 2 simply reduces to the well-known Bernstein–Vazirani algorithm [8].

5 Quantum sample complexity upper bounds

This section contains the description of two procedures for solving the task of learning an unknown Boolean linear function from quantum examples w.r.t. a product distribution. (Here, we assume perfect quantum examples, noisy examples will be taken into consideration in the next section.) It is subdivided into an approach which is applicable for arbitrary (albeit not full) bias in the product distribution and a strategy which produces better results but is only valid for small bias.

5.1 Arbitrary bias

As in the case of learning w.r.t. the uniform distribution, we intend to run the generalized Bernstein–Vazirani algorithm multiple times as a subroutine and then use our knowledge of the outcome of the subroutine together with probability-theoretic arguments. The main difficulty compared to the case of an example state arising from the uniform distribution lies in the fact that whereas an observation of $j_{n+1} = 1$ when performing the standard Bernstein–Vazirani algorithm guarantees that $j_1 \dots j_n$ equals the desired string, this is not true in the μ-biased case. Hence, we have to develop a different procedure of learning from the outcomes of the subroutine. For this purpose, we propose Algorithm 3.

Algorithm 3 Amplified Generalized Bernstein–Vazirani algorithm - Version 1

Input: m copies of $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$ for $a \in \{0,1\}^n$, where the number of copies is $m \geq C\left(\left\lceil \left(2\ln\left(\frac{1}{1-c+\frac{c^2}{2}}\right)\right)^{-1}\left(\ln(n)+\ln(\frac{2}{\delta})\right)\right\rceil\right)$ for a suitable constant $C > 0$, and $\mu \in (-1,1)^n$ and $c \in (0,1]$ s.t. D_μ is c-bounded.
Output: $a \in \{0,1\}^n$
Success Probability: $\geq 1-\delta$

1: **for** $1 \leq l \leq m$ **do**
2: Run Algorithm 2 on the lth copy of $|\psi_a\rangle$, store the output as $o^{(l)}$.
3: **end for**
4: **if** $\exists 1 \leq l \leq m : o^{(l)} \neq \perp$ **then**
5: **for** $1 \leq i \leq n$ **do**
6: Let $o_i := \max_{l:o^{(l)}\neq\perp} o_i^{(l)}$.
7: **end for**
8: Output $o = o_1 \dots o_n$.
9: **else if** $\forall 1 \leq l \leq m : o^{(l)} = \perp$ **then**
10: Output $o = \perp$.
11: **end if**

The amplification procedure in Algorithm 3 differs from the majority vote in the standard Bernstein–Vazirani learning procedure (w.r.t. the uniform distribution) as used in [11,13] in the following two ways: Instead of working on the level of the whole string, we use a componentwise strategy. And instead of taking a majority

vote over observed values, we take a maximum to account for the asymmetry in the probability of an observation error (see Theorem 2).

We now show that the number of copies postulated in Algorithm 3 is actually sufficient to achieve the desired success probability.

Theorem 3 *Let* $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$, $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ *s.t.* D_μ *is c-bounded for some* $c \in (0,1]$. *Then*

$$\mathcal{O}\left(\left(2\ln\left(\frac{1}{1-c+\frac{c^2}{2}}\right)\right)^{-1}\left(\ln(n)+\ln(\frac{2}{\delta})\right)\right)$$

copies of the quantum example state $|\psi_a\rangle$ *are sufficient to guarantee that, with probability* $\geq 1-\delta$, *Algorithm 3 outputs the string a.*

Proof We want to show that $\mathbb{P}[\text{Algorithm 3 does not output } a] \leq \delta$. We do so by treating separately the cases in which Algorithm 3 does not output a.
The first such case occurs if $o = \perp$. The second such case would be that there exists $1 \leq i \leq n$ s.t. $a_i = 0 \neq o_i$, but due to Theorem 2, this is an event of probability 0. The third and last such case is that there exists $1 \leq i \leq n$ s.t. $a_i = 1 \neq o_i$. Hence, we can decompose the probability of Algorithm 3 producing a wrong output as

$$\begin{aligned}&\mathbb{P}[\text{Algorithm 3 does not output } a]\\ &\quad = \mathbb{P}[\text{Algorithm 3 outputs } \perp] + \mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \neq o_i]. \end{aligned} \tag{5.1}$$

First, we bound the probability of the algorithm outputting $\perp$ (i.e., of each subroutine failing) as follows:

$$\begin{aligned}&\mathbb{P}[\text{Algorithm 3 outputs } \perp]\\ &\quad = \mathbb{P}[\forall 1 \leq l \leq m : \text{Algorithm 2 applied to } |\psi_a\rangle \text{ outputs } \perp]\\ &\quad = \left(\frac{1}{2}\right)^m,\end{aligned}$$

where the last step uses Theorem 2 and that the training data consists of independent copies of $|\psi_a\rangle$, i.e., is given as a product state. The choice of m now guarantees that this last term is $\leq \frac{\delta}{2}$ (if we choose the constant $C > 0$ sufficiently large).

Now we bound the second term in Eq. (5.1). We make the following observation: Suppose $1 \leq i \leq n$ is s.t. $a_i = 1$. As the Fourier coefficients, and with them the output probabilities, factorize, the probability of Algorithm 2 outputting a string $j_1 \ldots j_n$ with $j_i = 1 = a_i$ is simply the probability of Algorithm 2 applied to only the subsystem state of $|\psi_a\rangle$ corresponding to the ith and the $(n+1)^{st}$ subsystem outputting a 1. By Theorem 2, this probability is

$$\mathbb{P}[j_i = 1] = \mathbb{P}[j_{n+1} = 1] \cdot \mathbb{P}[j_i = 1 | j_{n+1} = 1] = \frac{1}{2} \cdot (1 - \mu_i^2).$$

Hence, assuming $a_i = 1$, the probability of not observing a 1 at the ith position in any of the m runs of Algorithm 2 is $\left(1 - \frac{1}{2} \cdot (1 - \mu_i^2)\right)^m = \left(\frac{1}{2}(1 + \mu_i^2)\right)^m$. By c-boundedness of the distribution D_μ we get

$$\left(\frac{1}{2}(1+\mu_i^2)\right)^m \leq \left(\frac{1}{2} + \frac{1}{2}(1-c)^2\right)^m = \left(1 - c + \frac{c^2}{2}\right)^m.$$

So using the union bound, we arrive at

$$\begin{aligned}
&\mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \neq o_i] \\
&\quad = \mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \text{ and in } m \text{ runs no 1 is observed at the } i^{th} \text{ entry}] \\
&\quad \leq \sum_{i=1}^{n} \mathbb{P}[a_i = 1 \text{ and in } m \text{ runs no 1 is observed at the } i^{th} \text{ entry}] \\
&\quad \leq n \cdot \left(1 - c + \frac{c^2}{2}\right)^m.
\end{aligned}$$

The choice of m guarantees that this last term is $\leq \frac{\delta}{2}$ (if we choose the constant $C > 0$ sufficiently large).

We now combine this with Eq. (5.1) and obtain

$$\mathbb{P}[\text{Algorithm 3 does not output } a] \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta,$$

which finishes the proof. □

Remark 1 We want to comment shortly on the dependence of the sample complexity bound on the c-boundedness constant by considering extreme cases. As $c \to 0$, i.e., we allow more and more strongly biased distributions, the sample complexity goes to infinity. This reflects the fact that in the case of a fully biased underlying product distribution, only a single bit of information about a can be extracted, so exactly learning the string a is (in general) not possible.
For $c = 1$, i.e., the case of no bias, we simply obtain that $\mathcal{O}\left(\left(\ln(n) + \ln(\frac{2}{\delta})\right)\right)$ copies of the quantum example state are sufficient. Note that this does not coincide with the bound obtained for the standard Bernstein–Vazirani procedure which is independent of n. (This can easily be shown using Lemma 1.)

This discrepancy is due to the difference in "amplification procedures." Namely, in Algorithm 3 we do not explicitly make use of the knowledge that, given $j_{n+1} = 1$, we know the probability of $j_1 \ldots j_n = a_1 \ldots a_n$ because, whereas for $\mu = 0$ this probability equals 1, for $\mu \neq 0$ it can become small. Hence, for $\mu \neq 0$ our algorithm introduces an additional procedure to deal with the uncertainty of $j_1 \ldots j_n$ even knowing j_{n+1} and we see in the proof that this yields the additional $\ln(n)$ term. In the next subsection, we describe a way to get rid of exactly that $\ln(n)$ term for "small" bias.

5.2 Small bias

In this subsection, we want to study the case in which (v) of Theorem 3 gives a good bound. Namely, throughout this subsection we will assume that the c-boundedness constant is s.t. $n(1-c)^2 < \frac{1}{2}$ or, equivalently, $c > 1 - \frac{1}{\sqrt{2n}}$. This assumption will allow us to apply a different procedure to learn from the output of Algorithm 2 and thus obtain a different bound on the sample complexity of the problem. Note, however, that this requirement becomes more restrictive with growing n and can in the limit $n \to \infty$ only be satisfied by $c = 1$, i.e., for the underlying distributions being uniform. Also, we will from now on refer to c as c-boundedness parameter because the name "constant" would hide the n-dependence.

Our procedure for the case of small bias is given in Algorithm 4.

Algorithm 4 Amplified Generalized Bernstein–Vazirani algorithm - Version 2

Input: m copies of $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$ for $a \in \{0,1\}^n$, where the number of copies is $m \geq C\left(\frac{4}{(1-2n(1-c)^2)^2} \ln\left(\frac{2}{\delta}\right)\right)$, as well as $\mu \in [-1,1]^n$ and $c \in (0,1]$ s.t. D_μ is c-bounded.

Output: $a \in \{0,1\}^n$

Success Probability: $\geq 1-\delta$

1: **for** $1 \leq l \leq m$ **do**
2: Run Algorithm 2 on the l^{th} copy of $|\psi_a\rangle$, store the output as $o^{(l)}$.
3: **end for**
4: **if** $\exists 1 \leq l \leq m : o^{(l)} \neq \perp$ **then**
5: **for** $1 \leq i \leq n$ **do**
6: Let $o_i = \arg\max_{r\in\{0,1\}} |\{1 \leq l \leq m | o_i^{(l)} = r\}|$.
7: **end for**
8: Output $o = o_1 \ldots o_n$.
9: **else if** $\forall 1 \leq l \leq m : o^{(l)} = \perp$ **then**
10: Output $o = \perp$.
11: **end if**

Theorem 4 *Let* $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$, $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ *s.t.* D_μ *is* c*-bounded for some* $c \in (0,1]$ *satisfying* $c > 1 - \frac{1}{\sqrt{2n}}$. *Then*

$$\mathcal{O}\left(\frac{1}{(1-2n(1-c)^2)^2} \ln\left(\frac{1}{\delta}\right)\right)$$

copies of the quantum example state $|\psi_a\rangle$ *are sufficient to guarantee that, with probability* $\geq 1-\delta$, *Algorithm 4 outputs the string* a.

Note that due to the required lower bound on c the sample complexity upper bound basically loses its n-dependence. This is different from the result of Theorem 3, where n explicitly entered the upper bound.

Proof By Theorem 2, we have $\mathbb{P}[j_{n+1} = 1] = \frac{1}{2}$. Hence, the probability of observing $j_{n+1} = 1$ in at most $k-1$ of the m runs of Algorithm 2 is given by

$$\sum_{l=0}^{k-1} \binom{m}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{m-i} = \mathbb{P}\left[\mathrm{Bin}(m, \frac{1}{2}) \geq m-k\right],$$

where Bin denotes a binomial distribution.

Next we assume $k \leq \frac{m}{2}$ (this will be justified later in the proof) and use Hoeffding's inequality (Lemma 1) to obtain

$$\begin{aligned}\mathbb{P}\left[\mathrm{Bin}\left(m, \frac{1}{2}\right) \geq m-k\right] &= \mathbb{P}\left[\mathrm{Bin}\left(m, \frac{1}{2}\right) - \frac{m}{2} \geq m-k-\frac{m}{2}\right] \\ &\leq \exp\left(-\frac{2\left(\frac{m}{2}-k\right)^2}{m}\right). \end{aligned} \tag{5.2}$$

We will now search for the number of observations of $j_{n+1} = 1$ which is required to guarantee that the majority string is correct with high probability. Assume that we observe $j_{n+1} = 1$ in k runs of Algorithm 2, $k \in 2\mathbb{N}$. (The latter assumption clearly does not significantly change the number of copies.) Using (v) from Theorem 2, we see that

$$\begin{aligned}\mathbb{P}[\exists 1 \leq i \leq n : a_i \neq o_i] &\leq \mathbb{P}[\exists 1 \leq i \leq n : a_i = 0 \neq o_i] \\ &\quad + \mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \neq o_i] \\ &\leq 0 + \sum_{l=\lceil \frac{k}{2} \rceil}^{k} \binom{k}{l} \cdot (1-n(1-c)^2)^{k-l} \cdot (n(1-c)^2)^l \\ &= \mathbb{P}\left[\mathrm{Bin}(k, n(1-c)^2) \geq \frac{k}{2}\right],\end{aligned}$$

where the second inequality uses that the majority string can only be wrong if in at least half of the runs where we observed $j_{n+1} = 1$ there was some error in the remaining string.

Next we use Hoeffding's inequality and obtain, using our assumption $n(1-c)^2 < \frac{1}{2}$, that

$$\begin{aligned}&\mathbb{P}\left[\mathrm{Bin}(k, n(1-c)^2) \geq \frac{k}{2}\right] \\ &= \mathbb{P}\left[\mathrm{Bin}(k, n(1-c)^2) - kn(1-c)^2 \geq \frac{k}{2} - kn(1-c)^2\right] \\ &\leq \exp\left(-k\frac{(1-2n(1-c)^2)^2}{2}\right).\end{aligned}$$

We now set this last expression $\leq \frac{\delta}{2}$ for $\delta \in (0, 1)$ and rearrange the inequality to

$$k \geq \frac{2}{(1-2n(1-c)^2)^2} \ln\left(\frac{2}{\delta}\right). \tag{5.3}$$

Combining Eqs. (5.3) and (5.2) we now require

$$\exp\left(-\frac{2\left(\frac{m}{2} - \frac{2}{(1-2n(1-c)^2)^2}\ln\left(\frac{2}{\delta}\right)\right)^2}{m}\right) \overset{!}{\leq} \frac{\delta}{2}.$$

Rearranging this inequality gives

$$m^2 - 2m\left(\left(\frac{1-2n(1-c)^2}{2}\right)^{-2} - 1\right)\ln\left(\frac{2}{\delta}\right) + \left(\frac{1-2n(1-c)^2}{2}\right)^{-4}\ln^2\left(\frac{2}{\delta}\right) \geq 0.$$

By finding the zeros of this quadratic function, we get to the sufficient sample size

$$m \geq \left(\left(\frac{1-2n(1-c)^2}{2}\right)^{-2} - 1\right)\ln\left(\frac{2}{\delta}\right) + \sqrt{\left(\left(\left(\frac{1-2n(1-c)^2}{2}\right)^{-2} - 1\right)\ln\left(\frac{2}{\delta}\right)\right)^2 - \left(\frac{1-2n(1-c)^2}{2}\right)^{-4}\ln^2\left(\frac{2}{\delta}\right)}.$$

This is in particular guaranteed if

$$m \geq \frac{4}{(1-2n(1-c)^2)^2} \ln\left(\frac{2}{\delta}\right).$$

Note that this lower bound in particular implies $m \geq 2k$, as required earlier in the proof. This proves the claim of the theorem thanks to the union bound. □

Morally speaking, Theorem 4 shows that for product distributions which are close enough to the uniform distribution the sample complexity upper bound is the same as for the unbiased case. We conjecture that there is an explicit noise threshold above which this sample complexity cannot be reached (see the discussion in Sect. 6), but have not yet succeeded in identifying such a critical value.

In this section, we have discussed the case of quantum training data that perfectly represents the target function in a superposition state. Similar results can be proved in the case of noisy quantum training data. As the reasoning is analogous to the one presented here, the details are deferred to "Appendix A."

6 Sample complexity lower bounds

After proving upper bounds on the number of required quantum examples by exhibiting explicit learning procedures in the previous section, we now study the converse

question of sample complexity lower bounds. We will prove both classical and quantum sample complexity lower bounds and then relate them to the above results. Our proof strategy follows a state-discrimination-based strategy from [3].

6.1 Classical sample complexity lower bounds

We first prove a sample complexity lower bound for the classical version of our learning problem that upon comparison with our obtained quantum sample complexity upper bounds shows the advantage of quantum examples over classical training data in this setting. Neither the result nor the proof strategy are new, but we include them for completeness.

Theorem 5 *Let $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ s.t. μ is c-bounded for some $c \in (0,1]$. Let $\mathcal{A}$ be a classical learning algorithm and let $m \in \mathbb{N}$ be such that upon input of m examples of the form $(x_i, f^{(a)}(x_i))$, with x_i drawn i.i.d. according to D_μ, with probability $\geq 1-\delta$ w.r.t. the choice of training data, $\mathcal{A}$ outputs the string a. Then $m \geq \Omega(n)$.*

Proof Let A be a random variable uniformly distributed on $\{0,1\}^n$. (A describes the underlying string from the initial perspective of the learner.) Let $B = (B_1, \ldots, B_m)$ be a random variable describing the training data corresponding to the underlying string. Our proof will have three main steps: First, we prove a lower bound on $I(A:B)$ from the learning requirement. Second, we observe that $I(A:B) \leq m \cdot I(A:B_1)$. And third, we prove an upper bound on $I(A:B_1)$. Then combining the three steps will lead to a lower bound on m.

We start with the mutual information lower bound. Let $h(B) \in \{0,1\}^n$ denote the random variable describing the output hypothesis of the algorithm $\mathcal{A}$ upon input of training data B. Let $Z = \mathbb{1}_{\{h(B)=A\}}$. By the learning requirement we have $\mathbb{P}[Z=1] \geq 1-\delta$ and thus $H(Z) \leq H(\delta)$. Therefore we obtain

$$\begin{aligned} I(A:B) &= H(A) - H(A|B) \\ &\geq H(A) - H(A|B,Z) - H(Z) \\ &= H(A) - \mathbb{P}[Z=1]H(A|B,Z=1) - \mathbb{P}[Z=0]H(A|B,Z=0) - H(Z) \\ &\geq n - \mathbb{P}[Z=1]\cdot 0 - \delta n - H(\delta) \\ &= (1-\delta)n - H(\delta) \\ &= \Omega(n). \end{aligned}$$

We now show that from m examples we can gather at most m times as much information as from a single example. Here we directly cite from [3]. Namely,

$$\begin{aligned} I(A:B) = H(B) - H(B|A) &= H(B) - \sum_{i=1}^{m} H(B_i|A) \\ &\leq \sum_{i=1}^{m} H(B_i) - H(B_i|A) = \sum_{i=1}^{m} I(A:B_i) = m \cdot I(A:B_1). \end{aligned}$$

Here, the second step uses independence of the B_i conditioned on A, the third step uses subadditivity of the Shannon entropy, and the final step uses that the distributions of (A, B_i) are the same for all $1 \leq i \leq m$.

We come to the upper bound on the mutual information. Write $B_1 = (X, L)$ for $X \in \{-1, 1\}^n$ and $L \in \{0, 1\}$, i.e., with probability $D_\mu(x)$ we have $(X, L) = (x, f^{(a)}(x))$. Note that $I(A : X) = 0$ because X and A are independent random variables. Also, $I(A : L|X = 1 \ldots 1) = 0$ because $f^{(a)}(1 \ldots 1) = 0 \; \forall a \in \{0, 1\}^n$, and for $x \in \{-1, 1\}^n \setminus \{1 \ldots 1\}$

$$\begin{aligned} I(A : L|X = x) &= I(A_{\{i|X_i=-1\}} : L|X = x) \\ &= H(A_{\{i|X_i=-1\}}|X = x) - H(A_{\{i|X_i=-1\}}|L, X = x) \\ &= |\{i|x_i = -1\}| - (|\{i|x_i = -1\}| - 1) \\ &= 1. \end{aligned}$$

Here, the first step is due to the fact that $f^{(a)}(x)$ does not depend on the entries a_j with $x_j = 1$, the third step follows because $A_{\{i|x_i=-1\}}$ is uniformly distributed on a set of size $2^{|\{i|x_i=-1\}|}$ and $f^{(a)}$ assigns the labels 0 and 1 to half of the elements of that set, respectively.
This now implies

$$\begin{aligned} I(A : B_1) &= I(A : X) + I(A : L|X) \\ &= 0 + \sum_{x \in \{-1,1\}^n} D_\mu(x) I(A : L|X = x) \\ &= 1. \end{aligned}$$

Here, the first step is due to the chain rule for mutual information and the last step simply uses the fact that D_μ defines a probability distribution.
Now we combine our upper and lower bounds on the mutual information and obtain

$$m \geq (1 - \delta)n - H(\delta) = \Omega(n),$$

as claimed. □

Remark 2 The result of Theorem 5 is intuitively clear: In order to identify the underlying string the learning algorithm has to learn n bits of information. However, a condition of the form $f^{(a)}(x) = l$ for $x \in \{0, 1\}^n, l \in \{0, 1\}$, takes away at most one degree of freedom from the initial space $\{0, 1\}^n$ for a and thus from such an equality the algorithm can extract at most 1 bit of information. So at least n examples will be required. This observation is thus neither new nor surprising. But we want to emphasize that this analysis works independently of the product structure of the underlying distribution D_μ.

If we compare the classical lower bound from Theorem 5 with our quantum upper bounds from Theorems 3 and 4 , we conclude that quantum examples allow us to strictly outperform the best possible classical algorithm w.r.t. the number of required examples.

6.2 Quantum sample complexity lower bounds

We can use a similar argument to prove quantum sample complexity lower bounds. Note that steps 1 and 2 carry over with (almost) no changes. Only the analysis of step 3 changes significantly. Even though this proof strategy is possible, as in [3] it can be improved upon by an argument based on state discrimination. We will thus follow this same approach.

An n-independent quantum sample complexity lower bound is given in the following

Lemma 5 *Let* $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$, $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ *s.t.* D_μ *is* c*-bounded for some* $c \in (0,1]$. *Let* $\mathcal{A}$ *be a quantum learning algorithm and let* $m \in \mathbb{N}$ *be such that upon input of* m *copies of* $|\psi_a\rangle$, *with probability* $\geq 1-\delta$, $\mathcal{A}$ *outputs the string* a. *Then* $m \geq \Omega(\frac{1}{c}\ln(\frac{1}{\delta}))$.

Remark 3 Note that any quantum sample complexity lower bound will also lower bound the classical sample complexity. Hence, Lemma 2 also holds in the scenario of the previous subsection, which is why we did not discuss the δ-dependence there.

Proof Let $a, b \in \{0,1\}^n$ s.t. there is exactly one $1 \leq i \leq n$ s.t. $a_i \neq b_i$. As $\mathcal{A}$ is able to distinguish the quantum states $|\psi_a\rangle^{\otimes m}$ and $|\psi_b\rangle^{\otimes m}$ with success probability $\geq 1-\delta$, we have $|\langle\psi_a|\psi_b\rangle^m| \leq 2\sqrt{\delta(1-\delta)}$ (see subsection 3.2). We compute

$$\begin{aligned}\langle\psi_a|\psi_b\rangle &= \sum_{x,y\in\{-1,1\}^n} \sqrt{D_\mu(x)D_\mu(y)}\langle x, f^{(a)}(x)|y, f^{(b)}(y)\rangle \\ &= \sum_{x\in\{-1,1\}^n} D_\mu(x)\delta_{f^{(a)}(x),f^{(b)}(x)}.\end{aligned}$$

By our assumption on a and b, $\delta_{f^{(a)}(x),f^{(b)}(x)} \geq \delta_{x_i,1}$. Therefore

$$\langle\psi_a|\psi_b\rangle \geq \mathbb{P}_{D_\mu}[x_i = 1] = \frac{1+\mu_i}{2}.$$

We now combine this with our upper bound and rearrange to obtain

$$\begin{aligned}m &\geq \left(\ln\left(\frac{1+\mu_i}{2}\right)\right)^{-1}\left(\ln(2) + \frac{1}{2}\ln(\delta(1-\delta))\right) \\ &\geq \Omega\left(\frac{1}{\mu_i - 1}\ln(\delta)\right) \\ &\geq \Omega\left(\frac{1}{c}\ln\left(\frac{1}{\delta}\right)\right),\end{aligned}$$

where we used the elementary inequality $\frac{1}{x-1} - \left(\ln\left(\frac{1+x}{2}\right)\right)^{-1} \geq 0$ for $x \in [0,1)$ combined with $\ln(\delta) \leq 0$. □

We will compare this lower bound with our upper bound(s) from Sect. 5 later on. Now we turn to the n-dependent part of the sample complexity lower bound.

Theorem 6 *Let* $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$, $a \in \{0,1\}^n$, *and* $\mu \in (-1,1)$ *be such that* $\mu_i = \mu \geq 1 - \frac{1}{\ln(n)}$ *for all* $1 \leq i \leq n$. *Let* $\mathcal{A}$ *be a quantum learning algorithm and let* $m \in \mathbb{N}$ *be such that upon input of* m *copies* $|\psi_a\rangle$, *with probability* $\geq 1-\delta$, $\mathcal{A}$ *outputs the string* a, *for* $0 < \delta \leq \frac{1}{3}$. *Then* $m \geq \Omega(\ln(n))$.

Before going into the detailed proof, we give an overview over its underlying idea. The learning assumption implies that $\mathcal{A}$ is able to identify a state from the ensemble $\mathcal{E} = \left\{\left(\frac{1}{2^n}, |\psi_a\rangle^{\otimes m}\right)\right\}_{a\in\{0,1\}^n}$ with success probability $\geq 1-\delta$. Thus we will obtain a lower bound on m by proving an upper bound on the optimal success probability for this state identification task.

Recall that by Theorem 1, the optimal success probability can be upper bounded by the square root of the PGM success probability. Moreover, by Lemma 3, the latter can be computed via the Gram matrix of the ensemble. Thus, we now first study the Gram matrix and its square root and then use these results to bound the optimal success probability.

We first recall a well-known result on the diagonalization of matrices with a specific structure, namely matrices whose entries can be written as Boolean function of the sum of the indices.

Lemma 6 *Let* $G \in \mathbb{R}^{2^n\times 2^n}$ *be a matrix with entries given by* $G(a,b) = g(a+b)$ *for* $a, b \in \{0,1\}^n$ *and a function* $g : \{0,1\}^n \to \mathbb{R}$. *Then*

$$(HGH^{-1})(a,b) = 2^n \hat{g}(a)\delta_{a,b},$$

with $H \in \mathbb{R}^{2^n\times 2^n}$ *given by* $H(a,b) = \frac{(-1)^{a\cdot b}}{\sqrt{2^n}}$. *In other words, the set of eigenvalues of* G *is given by* $\{2^n\hat{g}(a) \mid a \in \{0,1\}^n\}$ *and* G *is unitarily diagonalized by* H.

Proof The proof can be found in [3], we reproduce it in "Appendix B" □

We will later apply this result for G being the Gram matrix corresponding to the ensemble in our state identification task. Motivated by Lemma 3, we first use the diagonalization of such a matrix to explicitly compute the diagonal entries of the matrix square root.

Corollary 1 *Let* $G \in \mathbb{R}^{2^n\times 2^n}$ *be a matrix with entries given by* $G(a,b) = g(a+b)$ *for* $a, b \in \{0,1\}^n$ *and a function* $g : \{0,1\}^n \to \mathbb{R}$. *Then, for every* $a \in \{0,1\}^n$

$$\sqrt{G}(a,a) = \frac{1}{\sqrt{2^n}} \sum_{j\in\{0,1\}^n} \sqrt{\hat{g}(j)}.$$

Proof The proof can be found in [3], we reproduce it in "Appendix B." □

With this, we can now prove Theorem 6:

Proof of Theorem 6 As discussed above, we consider the problem of state identification with the ensemble $\mathcal{E} = \left\{\left(\frac{1}{2^n}, |\psi_a\rangle^{\otimes m}\right)\right\}_{a\in\{0,1\}^n}$. By Lemma 3, with the Gram matrix $G_m(a,b) := \frac{1}{2^n}\langle\psi_a|\psi_b\rangle^m$ we can write the success probability as

$$P^{PGM}(\mathcal{E}) = \sum_{a\in\{0,1\}^n} \sqrt{G_m}(a,a)^2.$$

In our scenario, the Gram matrix has entries

$$\begin{aligned} G_m(a,b) &= \frac{1}{2^n}\langle\psi_a|\psi_b\rangle^m \\ &= \frac{1}{2^{n+m}}\left(1+\mu^{\mathrm{d}_H(a,b)}\right)^m = \frac{1}{2^{n+m}}\left(1+\mu^{\mathrm{d}_H(a+b,0)}\right)^m. \end{aligned}$$

This can, e.g., be shown by induction on n when observing that

$$\begin{aligned} &\mathbb{P}_{D_\mu}[f^{(a)}(x) = f^{(b)}(x)] \\ &= \mathbb{P}_{D_\mu}\left[f^{(a_{1:n-1})}(x_{1:n-1}) = f^{(b_{1:n-1})}(x_{1:n-1}) \ \wedge\ a_n\frac{1-x_n}{2} = b_n\frac{1-x_n}{2}\right] \\ &\quad + \mathbb{P}_{D_\mu}\left[f^{(a_{1:n-1})}(x_{1:n-1}) \neq f^{(b_{1:n-1})}(x_{1:n-1}) \ \wedge\ a_n\frac{1-x_n}{2} \neq b_n\frac{1-x_n}{2}\right]. \end{aligned}$$

In particular, we can write $G_m(a,b) = f_m(a+b)$ for the function $f_m(x) = \frac{1}{2^{n+m}}\left(1+\mu^{\mathrm{d}_H(x,0)}\right)^m$. From now on, we will write $|x| := \mathrm{d}_H(x,0)$. By Corollary 1, we can upper bound the diagonal entries of $\sqrt{G_m}$ (and thus the PGM and the optimal success probability) by upper bounding the (unbiased) Fourier coefficients of f_m. To this end, consider for $j \in \{0,1\}^n$

$$\begin{aligned} 0 \leq \hat{f}_m(j) &= \mathbb{E}_{z\sim U(\{0,1\}^n)}\left[\frac{1}{2^{n+m}}\left(1+\mu^{|z|}\right)^m(-1)^{j\cdot z}\right] \\ &= \frac{1}{2^{n+m}}\sum_{L=0}^{m}\binom{m}{L}\mathbb{E}_{z\sim U(\{0,1\}^n)}\left[\mu^{L|z|}(-1)^{j\cdot z}\right]. \end{aligned}$$

We now rewrite the expectations on the right-hand side

$$\begin{aligned} &\mathbb{E}_{z\sim U(\{0,1\}^n)}\left[\mu^{L|z|}(-1)^{j\cdot z}\right] \\ &= \frac{1}{2^n}\sum_{\ell=0}^{n}\sum_{k=\max\{0,\ell-(n-|j|)\}}^{\min\{\ell,|j|\}}\binom{|j|}{k}\binom{n-|j|}{\ell-k}(-1)^k\mu^{L\cdot\ell} \\ &= \frac{1}{2^n}\sum_{k=0}^{|j|}\binom{|j|}{k}(-1)^k\sum_{\ell=k}^{k+n-|j|}\binom{n-|j|}{\ell-k}\mu^{L\cdot\ell} \end{aligned}$$

$$= \frac{1}{2^n} \sum_{k=0}^{|j|} \binom{|j|}{k} (-1)^k \mu^{L\cdot k} \underbrace{\sum_{\ell=0}^{n-|j|} \binom{n-|j|}{\ell} \mu^{L\cdot \ell}}_{=(1+\mu^L)^{n-|j|}}$$

$$= \frac{\left(1+\mu^L\right)^{n-|j|}}{2^n} \underbrace{\sum_{k=0}^{|j|} \binom{|j|}{k} (-1)^k \mu^{L\cdot k}}_{=(1-\mu^L)^{|j|}}$$

$$= \frac{\left(1+\mu^L\right)^{n-|j|} \left(1-\mu^L\right)^{|j|}}{2^n}.$$

This allows us to upper bound the Fourier coefficients of f as follows:

$$\begin{aligned}
\hat{f}_m(j) &= \frac{1}{2^{n+m}} \sum_{L=0}^{m} \binom{m}{L} \left(\frac{1+\mu^L}{2}\right)^{n-|j|} \left(\frac{1-\mu^L}{2}\right)^{|j|} \\
&\leq \frac{1}{2^{n+m}} \sum_{L=0}^{m} \binom{m}{L} \left(\frac{1+\mu}{2}\right)^{n-|j|} \left(\frac{1-\mu^m}{2}\right)^{|j|} \\
&= \frac{1}{2^n} \left(\frac{1+\mu}{2}\right)^{n-|j|} \left(\frac{1-\mu^m}{2}\right)^{|j|}.
\end{aligned}$$

According to Lemma 6, this now gives us the following upper bound on the diagonal entries of the root of the Gram matrix

$$\begin{aligned}
\sqrt{G_m}(a,a) &\leq \frac{1}{2^n} \sum_{j\in\{0,1\}^n} \sqrt{\left(\frac{1+\mu}{2}\right)^{n-|j|} \left(\frac{1-\mu^m}{2}\right)^{|j|}} \\
&= \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \sqrt{\left(\frac{1+\mu}{2}\right)^{n-k} \left(\frac{1-\mu^m}{2}\right)^{k}} \\
&= \frac{1}{2^n} \left(\sqrt{\frac{1+\mu}{2}} + \sqrt{\frac{1-\mu^m}{2}}\right)^n,
\end{aligned}$$

and this in turn allows us to bound the PGM success probability as

$$\begin{aligned}
P^{PGM}(\mathcal{E}) &= \sum_{a\in\{0,1\}^n} \sqrt{G_m}(a,a)^2 \\
&\leq \frac{1}{2^n} \left(\sqrt{\frac{1+\mu}{2}} + \sqrt{\frac{1-\mu^m}{2}}\right)^{2n} \\
&= \left(\frac{1}{2}\left(\sqrt{1+\mu} + \sqrt{1-\mu^m}\right)\right)^{2n}.
\end{aligned}$$

We combine this with our learning requirement and Theorem 1 to obtain

$$1-\delta \leq P^{opt}(\mathcal{E}) \leq \sqrt{P^{PGM}(\mathcal{E})} \leq \left(\frac{1}{2}\left(\sqrt{1+\mu}+\sqrt{1-\mu^m}\right)\right)^n.$$

This can be rearranged (using $\delta < \frac{1}{3}$) to

$$m = \frac{-\log\left(1-\left(2\cdot\sqrt[n]{1-\delta}-\sqrt{1+\mu}\right)^2\right)}{\log\frac{1}{\mu}}.$$

With $\log(1+x) \leq x$ we obtain $\frac{1}{\log\frac{1}{\mu}} \geq \frac{1}{\frac{1}{\mu}-1} = \frac{\mu}{1-\mu}$ and

$$-\log\left(1-\left(2\cdot\sqrt[n]{1-\delta}-\sqrt{1+\mu}\right)^2\right) \geq \left(2\cdot\sqrt[n]{1-\delta}-\sqrt{1+\mu}\right)^2.$$

For $\mu \geq 1-\frac{1}{\ln(n)}$ we now obtain (for n large enough)

$$m \geq (\ln(n)-1)\cdot\left(2\sqrt{\frac{2}{3}}-\sqrt{2}\right) = \Omega\left(\ln(n)\right),$$

and this finishes the proof. □

Note that this proof strategy also yields for a strictly increasing function $g : \mathbb{N} \to \mathbb{R}_{>0}$ with $\lim_{n\to\infty} g(n) = \infty$ and for a distribution D_μ with $\mu_i \geq 1-\frac{1}{g(n)}$ for all $1 \leq i \leq n$ the sample complexity lower bound $\Omega(g(n))$ (for n large enough). This is consistent with the intuition that solving the learner problem becomes harder when the distribution is more strongly biased towards the uninformative instance with all entries equal to 1.

We now compare this lower bound to our previously obtained upper bounds. First, we consider the n-independent part of the bounds. When comparing Theorem 3 with Lemma 5, we obtain

$$\Omega\left(\frac{1}{c}\ln\left(\frac{1}{\delta}\right)\right) \leq m \leq \mathcal{O}\left(\left(\ln\left(\frac{1}{1-c+\frac{c^2}{2}}\right)\right)^{-1}\ln\left(\frac{1}{\delta}\right)\right).$$

We study this for $\delta \ll 1$ (high confidence) and $c \ll 1$ (high bias). Then Taylor expansion shows

$$\left(\ln\left(\frac{1}{1-c+\frac{c^2}{2}}\right)\right)^{-1} = \frac{1}{c}+\frac{c}{6}+\mathcal{O}(c^2) \quad \text{for } c \ll 1.$$

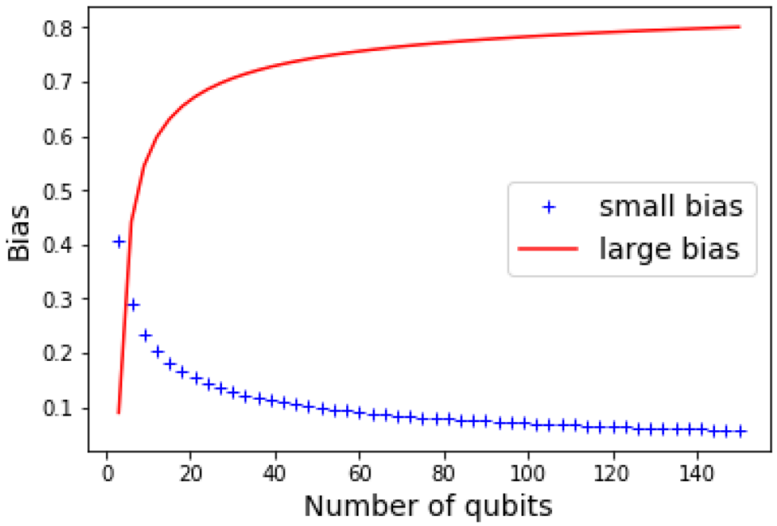


Fig. 1 A plot comparing the maximal bias allowed in Theorem 4 (depicted by the blue crosses) with the minimal bias required in Theorem 6 (depicted by the red line) (Color figure online)

Hence, lower and upper bounds coincide in the relevant region for δ and c, so the n-independent part of the sample complexity upper bound provided by Algorithm 3 is optimal.

However, in comparing Theorem 4 with Lemma 5 we see a discrepancy between lower and upper bound for the relevant region $\delta \ll 1$ and $c-(1-\frac{1}{\sqrt{2n}}) \ll 1$. Therefore we conjecture that the c-dependence of the upper bound arising from Theorem 4 is not optimal.

Now we compare the bounds w.r.t. the n-dependence, i.e., we compare Theorem 3 with Theorem 6, and obtain

$$\Omega\left(\ln(n)\right) \leq m \leq \mathcal{O}\left(\frac{1}{c}\ln(n)\right).$$

But in Theorem 6, we assumed that $\mu_i \geq 1-\frac{1}{\ln(n)}$ for all $1 \leq i \leq n$. When considering values for μ lying on this threshold, we can rephrase this as condition on the (then n-dependent) c-boundedness parameter, namely $c \leq \frac{1}{\ln(n)}$. So when honestly including the n-dependence of c, our comparison becomes

$$\Omega\left(\ln(n)\right) \leq m \leq \mathcal{O}\left(\ln^2(n)\right)$$

and is thus not tight.

Finally, we want to point towards a second unsatisfactory aspect of our results. We provide an n-dependent quantum sample complexity lower bound for "large" noise and an n-independent quantum sample complexity upper bound for "small" noise. However, there is a large discrepancy between the obtained characterizations of "small" and "large" noise. That this already becomes relevant for moderate n can be seen in Fig. 1.

Hence, we did not succeed in identifying a bias threshold beyond which the sample complexity qualitatively differs from the unbiased case, but merely provided a region in

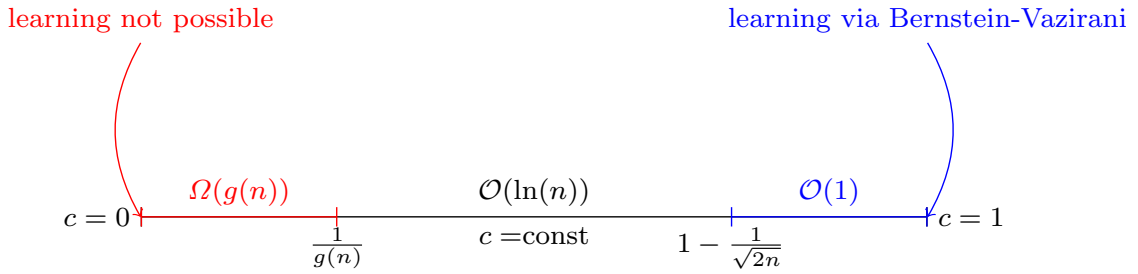


Fig. 2 Overview of the quantum sample complexity upper and lower bounds from Theorems 3, 4 and 6 depending on the c-boundedness parameter (without noise in the training data). Here, $g : \mathbb{N} \to \mathbb{R}_{>0}$ is a strictly increasing function with $\lim_{n\to\infty} g(n) = \infty$ (Color figure online)

which such a threshold would lie. To improve upon our results, it would be necessary to modify either the proof of Theorem 4 to allow for stronger bias or the proof of Theorem 6 to allow for weaker bias. In particular, it would be interesting to obtain a non-trivial quantum sample complexity lower bound for constant bias, i.e., without introducing n-dependence into the c-boundedness parameter. However, we currently do not see whether our proof strategies admit such an improvement.

7 Conclusion and outlook

In this paper, we extended a well-known quantum learning strategy for linear functions from the uniform distribution to biased product distributions. This approach naturally led to a distinction between a procedure for arbitrary (not full) bias and a procedure for small bias, the latter with a significantly better performance. Moreover, we showed that the second procedure is (to a certain degree) stable w.r.t. noise in the training data and in the performed quantum gates. Finally, we also provided lower bounds on the size of the training data required for the learning problem, both in the classical and in the quantum setting. The sample complexity upper and lower bounds in the case of no noise are summarized in Fig. 2.

We want to conclude by outlining some open questions for future work:

- Can we identify a bias threshold s.t. the optimal sample complexity below the threshold differs qualitatively from the one above it?
- Is our learning procedure for small bias also stable w.r.t. different types of noise in the training data, e.g., malicious noise?
- Our explicit learning algorithms also give upper bounds on the computational complexity of our learning problem. Can we find corresponding lower bounds to facilitate a discussion of optimality w.r.t. runtime?
- Can we find more examples of learning tasks (i.e., function classes) where quantum training data yields an advantage w.r.t. sample and/or time complexity?

Acknowledgements Open Access funding provided by Projekt DEAL. First, I want to thank my supervisor Michael Wolf for several stimulating discussions concerning questions of quantum learning. Also, I want to thank Benedikt Graswald for proofreading a first draft of this paper and for his constructive comments. Finally, I am grateful to Andrea Rocchetto for useful comments to improve the result of "Appendix A.3" and for suggesting further references. Also, I thank the reviewers for their constructive criticism. Support

from the TopMath Graduate Center of the TUM Graduate School at the Technische Universität München, Germany, and from the TopMath Program at the Elite Network of Bavaria is gratefully acknowledged.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

Appendix

A Stability w.r.t. noise

Both algorithms presented in Sect. 5 implicitly assume that the quantum example state perfectly represents the underlying function and that all quantum gates performed during the computation are perfectly accurate. In this section, we relax these assumptions. We will do so separately, but our analysis shows that moderate noise in the training data and moderately faulty quantum gates can be tolerated at the same time.

A.1 Noisy training data

One of the most well-studied noise models in classical learning theory is that of random classification noise. Here, the training data are assumed to be s.t. with probability $1-\eta$, the learning algorithm obtains a correct example, and with probability η, the examples label is flipped. In [4], this is translated to a quantum example state which in our notation has the form

$$|\varphi_a^{\text{noisy}}\rangle = \sqrt{1-\eta}\left(\sum_{x\in\{-1,1\}}\sqrt{D_\mu(x)}|x,\, f^{(a)}(x)\rangle\right) + \sqrt{\eta}\left(\sum_{x\in\{-1,1\}}\sqrt{D_\mu(x)}|x,\, f^{(a)}(x)\oplus 1\rangle\right).$$

We will only shortly comment on how to battle this type of noise with our learning strategy at the end of this subsection. Instead, our focus will be on a performance analysis of our algorithm in the case of noisy training data similar to [13]. This means that we now assume our quantum example state to be of the form

$$|\psi_a^{\text{noisy}}\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, \sum_{i=1}^n a_i \frac{1-x_i}{2} + \xi_{x_i}^i\rangle,$$

where the $\xi_{x_i}^i$, for $1 \leq i \leq n$ and $x_i \in \{-1, 1\}$, are independent random variables distributed according to Bernoulli distributions with parameters η^i (i.e., $\mathbb{P}[\xi_{x_i}^i = 1] = \eta^i = 1 - \mathbb{P}[\xi_{x_i}^i = 0]$ for all $1 \leq i \leq n$) and addition is understood modulo 2.

Here, we choose a noise model that is rather general but we make an important restriction. Namely, we do not allow a noise ξ_x that depends in an arbitrary way on x but rather we require the noise to have a specific sum structure $\xi_x = \sum_{i=1}^n \xi_{x_i}^i$. This requirement will later imply that also the noisy Fourier coefficients factorize. As this factorization is crucial for our analysis, with our strategy we cannot generalize the results of [13] on that more general noise model.

We first examine the result of applying the same procedure as in Algorithm 2 to a copy of a noisy quantum example state $|\psi_a^{\text{noisy}}\rangle$. To simplify referencing, we write this down one more time as Algorithm 5 even though the procedure is exactly the same, only the form of the input changes.

Algorithm 5 Generalized Bernstein–Vazirani algorithm with noisy training data

Input: $|\psi_a^{\text{noisy}}\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, \sum_{i=1}^n a_i \frac{1-x_i}{2} + \xi_{x_i}^i\rangle$, as well as $\mu \in [-1, 1]$
Output: See Theorem 7
Success Probability: $\frac{1}{2}$.

1: Perform the μ-biased QFT H_μ on the first n qubits, obtain the state $(H_\mu \otimes \mathbb{1})|\psi_a^{\text{noisy}}\rangle$.
2: Perform a Hadamard gate on the last qubit, obtain the state $(H_\mu \otimes H)|\psi_a^{\text{noisy}}\rangle$.
3: Measure each qubit in the computational basis and observe outcome $j = j_1 \dots j_{n+1}$.
4: **if** $j_{n+1} = 0$ **then** ▷ This corresponds to a failure of the algorithm.
5: Output $o = \perp$.
6: **else if** $j_{n+1} = 1$ **then** ▷ This corresponds to a success of the algorithm.
7: Output $o = j_1 \dots j_n$.
8: **end if**

Similarly to our previous analysis, we will first study the Fourier coefficients that are relevant for the sampling process in Algorithm 5.

Lemma 7 *Let $a \in \{0, 1\}^n$. Let $\xi_{x_i}^i$, for $1 \leq i \leq n$ and $x_i \in \{-1, 1\}$, be independent Bernoulli distributions, let $g^{(a)}(x) := (-1)^{\sum_{i=1}^n a_i \frac{1-x_i}{2} + \xi_{x_i}^i}$ and let $\mu \in (-1, 1)$. Then the μ-biased Fourier coefficients of $g^{(a)}$ satisfy: For $y \in \{0, 1\}^n$, with probability*

$$\prod_{l=1}^n \left(y_l \cdot 2\eta^l(1-\eta^l) + (1-y_l)\cdot(1-2\eta^l(1-\eta^l)) \right),$$

it holds that

$$\hat{g}_{\mu}^{(a)}(j) = \prod_{l:a_l=0} \left(y_l \cdot (-1)^{b_l} \left((1-j_l)\mu_l + j_l\sqrt{1-\mu_l^2} \right) + (1-y_l) \cdot (-1)^{b_l}(1-j_l) \right)$$
$$\cdot \prod_{l:a_l=1} \left(y_l \cdot (-1)^{b_l}(1-j_l) + (1-y_l) \cdot (-1)^{b_l} \left((1-j_l)\mu_l + j_l\sqrt{1-\mu_l^2} \right) \right).$$

Proof The proof is analogous to the one of Lemma 4, see "Appendix B." □

We now make a step analogous to the one from Lemma 4 to Theorem 2 in order to understand the output of Algorithm 5.

Theorem 7 *Let* $|\psi_a^{noisy}\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, \sum_{i=1}^n a_i \frac{1-x_i}{2} + \xi_{x_i}^i\rangle$ *be a noisy quantum example state,* $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$. *Then Algorithm 5 provides an outcome* $|j_1 \dots j_{n+1}\rangle$ *with the following properties:*

(i) $\mathbb{P}[j_{n+1} = 0] = \frac{1}{2} = \mathbb{P}[j_{n+1} = 1]$.

(ii) *For any* $1 \le i \le n$, *with probability* $1 - 2\eta^i(1-\eta^i)$ *it holds that*

$$\mathbb{P}[a_i = 0 \neq j_i | j_{n+1} = 1] = 0, \quad \mathbb{P}[a_i = 1 \neq j_i | j_{n+1} = 1] = \mu^2.$$

(iii) *For any* $1 \le i \le n$, *with probability* $2\eta^i(1-\eta^i)$ *it holds that*

$$\mathbb{P}[a_i = 0 \neq j_i | j_{n+1} = 1] = 1 - \mu^2, \quad \mathbb{P}[a_i = 1 \neq j_i | j_{n+1} = 1] = 1.$$

Note that in the scenario of Theorem 7 the underlying distribution D_μ is known to the algorithm as μ is provided as part of the input (see Algorithm 5). Building on this subroutine, we will now describe an amplified procedure for moderate noise (which is made precise in Theorem 8) in Algorithm 6 analogous to the one described in Sect. 5.2. Again, only the input changes, but we write the procedure down explicitly to simplify referencing.

Theorem 8 *Let* $|\psi_a^{noisy}\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, \sum_{i=1}^n a_i \frac{1-x_i}{2} + \xi_{x_i}^i\rangle$, *with* $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ *s.t.* D_μ *is c-bounded for some* $c \in (0,1]$ *satisfying* $c > 1 - \frac{1}{2\sqrt{n}}$. *Further assume that* $2\eta^i(1-\eta^i) < \frac{1}{5n}$ *for all* $1 \le i \le n$, *write* $\rho := \max_{1\le i\le n} 2\eta^i(1-\eta^i)$. *Then* $\mathcal{O}\left(\max\left\{\frac{1}{(1-5n\rho)^2}, \frac{1}{(1-4n(1-c)^2)^2}\right\} \ln\left(\frac{1}{\delta}\right)\right)$ *copies of the quantum example state* $|\psi_a\rangle$ *suffice to guarantee that with probability* $\ge 1-\delta$ *Algorithm 6 outputs the string a.*

As in Theorem 4, our restrictions on both the c-boundedness parameter and the noise strength lead to a basically n-independent sample complexity upper bound.

Proof The proof is analogous to the one of Theorem 4, see "Appendix B." □

Algorithm 6 Amplified Generalized Bernstein–Vazirani algorithm with noisy training data

Input: m copies of $|\psi_a^{\text{noisy}}\rangle = \sum_{x\in\{-1,1\}^n} \sqrt{D_\mu(x)}|x, \sum_{i=1}^{n} a_i \frac{1-x_i}{2} + \xi^i_{x_i}\rangle$ for $a \in \{0,1\}^n$, where the number of copies is $m \geq C\left(\max\left\{\frac{1}{(1-5n\rho)^2}, \frac{1}{(1-4n(1-c)^2)^2}\right\}\ln\left(\frac{1}{\delta}\right)\right)$, as well as $\mu \in [-1,1]^n$ and $c \in (0,1]$ s.t. D_μ is c-bounded.
Output: $a \in \{0,1\}^n$
Success Probability: $\geq 1-\delta$

1: **for** $1 \leq l \leq m$ **do**
2: Run Algorithm 5 on the l^{th} copy of $|\psi_a^{\text{noisy}}\rangle$, store the output as $o^{(l)}$.
3: **end for**
4: **if** $\exists 1 \leq l \leq m : o^{(l)} \neq \perp$ **then**
5: **for** $1 \leq i \leq n$ **do**
6: Let $o_i = \arg\max_{r\in\{0,1\}} |\{1 \leq l \leq m | o_i^{(l)} = r\}|$.
7: **end for**
8: Output $o = o_1 \ldots o_n$.
9: **else if** $\forall 1 \leq l \leq m : o^{(l)} = \perp$ **then**
10: Output $o = \perp$.
11: **end if**

The previous Theorem shows that if the bias is not too strong and if the noise is not too random (i.e., the probability of adding a random 1 is either very low or very high), then learning is possible with essentially the same sample complexity as in the case without noise (compare Theorem 4).
Note that the proof of Theorem 8 shows that the exact choices of the bounds (in our formulation $c > 1 - \frac{1}{2\sqrt{n}}$ and $2\eta^i(1-\eta^i) < \frac{1}{5n}$) are flexible to some degree with a trade-off. If we have a better bound on c, we can loosen our requirement on the η^i and vice versa.

Also observe that the requirement of "not too random noise" is natural. If $2\eta^i(1-\eta^i) \to \frac{1}{2}$ or, equivalently, $\eta^i \to \frac{1}{2}$, then the label in the noisy quantum example state becomes completely random and thus no information on the string a can be extracted from it. Our bound gives a quantitative version of this intuition.

Nevertheless, the restriction which we put on the noise can be considered quite strong because of its n-dependence. This can, however, be relaxed at the cost of a looser sample complexity upper bound. Namely, similarly to the difference between the proofs of Theorems 3 and 4 , if we, e.g., only assume $2\eta^i(1-\eta^i) < \frac{1}{5}$ for all $1 \leq i \leq n$, we can first for each coordinate separately bound the probability of the noise variables becoming relevant in at least $\frac{k}{5}$ runs using Hoeffding's inequality and then use the union bound. This will yield a quantum sample complexity upper bound with an n-dependent term of the form $\ln(n)$. Hence, if we assume a c-boundedness parameter strongly restricted as in Theorems 4 or 8, but obtain faulty training data states without an n-dependent noise bound as in Theorem 8, then we can still obtain a sample complexity upper bound with the same n-dependence as in Theorem 3.

Finally, as promised at the beginning of this subsection, we shortly describe how to use the ideas presented in this subsection in the case of random classification noise as

in [4]. If the quantum learning algorithm has access to copies of a quantum example state

$$|\varphi_a^{\text{noisy}}\rangle = \sqrt{1-\eta}\left(\sum_{x\in\{-1,1\}}\sqrt{D_\mu(x)}|x,\, f^{(a)}(x)\rangle\right) + \sqrt{\eta}\left(\sum_{x\in\{-1,1\}}\sqrt{D_\mu(x)}|x,\, f^{(a)}(x)\oplus 1\rangle\right),$$

then we observe that applying the μ-biased Fourier transform to SSthe first n qubits and the standard Fourier transform to the last qubit gives

$$(H_\mu^{\otimes n}\otimes H)\left(|\varphi_a^{\text{noisy}}\rangle\right) = \frac{\sqrt{1-\eta}+\sqrt{\eta}}{\sqrt{2}}|0,\ldots,0\rangle + \frac{\sqrt{1-\eta}-\sqrt{\eta}}{\sqrt{2}}\sum_{j\in\{0,1\}}\hat{g}_\mu(j)|j,1\rangle.$$

Hence, compared to the scenario studied in section 5 the probabilities of observing a certain string as measurement outcome are simply scaled by a factor of $(\sqrt{1-\eta}\pm\sqrt{\eta})^2 = 1\pm 2\sqrt{\eta(1-\eta)}$. So our analysis carries over almost directly. We do not give the detailed reasoning here but only mention that incorporating the now rescaled probabilities basically changes the sample complexity upper bounds from the non-noisy case by a factor of $\frac{1}{(\eta-\frac{1}{2})^2}$, which is again in accordance with the intuition that the learning task becomes hard—and eventually impossible—for $\eta\to\frac{1}{2}$.

A.2 Faulty quantum gates

We now turn to the (more realistic) setting where the quantum gates in our computation (i.e., the μ-biased quantum Fourier transforms) are not implemented exactly but only approximately. In this scenario, we obtain

Lemma 8 *Let $|\psi_a\rangle = \sum_{x\in\{-1,1\}^n}\sqrt{D_\mu(x)}|x,\, f^{(a)}(x)\rangle$ be a quantum example state, with $a\in\{0,1\}^n$, $\mu\in(-1,1)^n$. Then a version of Algorithm 2 with H_μ replaced by $H_{\tilde{\mu}}$ for $\|H_\mu - H_{\tilde{\mu}}\|_2 \le \varepsilon$ provides an outcome $|j_1\ldots j_{n+1}\rangle$ with the following properties:*

(i) $|\mathbb{P}[j_{n+1}=0]-\frac{1}{2}|\le\varepsilon$ *and* $|\mathbb{P}[j_{n+1}=1]-\frac{1}{2}|\le\varepsilon$,
(ii) $|\mathbb{P}[j_1\ldots j_n = a\,|\,j_{n+1}=1] - \prod_{l:a_l=1}(1-\mu_l^2)|\le\varepsilon$,
(iii) *for $c\neq a$:*

$$\left|\mathbb{P}[j_1\ldots j_n = c\,|\,j_{n+1}=1] - \prod_{l:a_l=0}(1-c_l)\cdot\prod_{l:a_l=1}\left((1-c_l)\mu_l^2 + c_l(1-\mu_l^2)\right)\right|\le\varepsilon,$$

(iv) $\mathbb{P}[\exists 1 \leq i \leq n : a_i = 0 \neq j_i | j_{n+1} = 1] \leq \varepsilon$, *and*

(v) $\mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \neq j_i | j_{n+1} = 1] \leq \sum_{i=1}^{n} \mu_i^2 + \varepsilon$. *In particular, if* D_μ *is* c*-bounded, then* $\mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \neq j_i | j_{n+1} = 1] \leq n(1-c)^2 + \varepsilon$.

Proof This follows from Theorem 2 because the outcome probabilities are the squares of the amplitudes, and thus, the difference in outcome probabilities can be bounded by the 2-norm of the difference of the quantum states after applying the biased quantum Fourier transform and its approximate version. □

Now we can proceed analogously to the proof strategy employed in Theorem 8 to derive

Theorem 9 *Let* $|\psi_a\rangle = \sum_{x \in \{-1,1\}^n} \sqrt{D_\mu(x)} |x, f^{(a)}(x)\rangle$, $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ *s.t.* D_μ *is* c*-bounded for some* $c \in (0,1]$ *satisfying* $c > 1 - \sqrt{\frac{1-2\varepsilon}{2n}}$. *Then*

$$\mathcal{O}\left(\max\left\{\frac{1}{(1-2\varepsilon)^2}, \frac{1}{1-2(n(1-c)^2+\varepsilon)^2}\right\} \ln\left(\frac{1}{\delta}\right) + \varepsilon\right)$$

copies of the quantum example state $|\psi_a\rangle$ *suffice to guarantee that, with probability* $\geq 1-\delta$, *a version of Algorithm 4 with* H_μ *replaced by* $H_{\tilde{\mu}}$ *for* $\left\|H_\mu - H_{\tilde{\mu}}\right\|_2 \leq \varepsilon \in (0, \frac{1}{2})$ *outputs the string* a.

In particular, the sample complexity upper bound from Theorem 4 remains basically untouched if quantum gates with small error are used.

A.3 The case of unknown underlying distributions

An interesting consequence of the result of the previous subsection is the possibility to drop the assumption of prior knowledge of the underlying product distribution, as was already observed in [17] for a similar scenario. The important observations towards this end are given in this subsection.

Lemma 9 (Lemma 5 in [17])

Let $A = A_n \cdots A_1$ *be a product of unitary operators* A_j. *Assume that for every* A_j *there exists an approximation* $\tilde{A}_j$ *s.t.* $\left\|A_j - \tilde{A}_j\right\| \leq \varepsilon_j$. *Then it holds that*

$$\left\|A_n \cdots A_1 - \tilde{A}_n \cdots \tilde{A}_1\right\| \leq \sum_{j=1}^{n} \varepsilon_j,$$

i.e., the operator $\tilde{A} := \tilde{A}_n \cdots \tilde{A}_1$ *is an* ε*-approximation to* A *w.r.t. the operator norm.*

Proof This can be proven by induction using the triangle inequality and the fact that a unitary operator has operator norm equal to 1. For details, the reader is referred to [17]. □

This can be used to derive (compare again [17])

Corollary 2 *Let $\mu \in (-1,1)^n$ be s.t. the distribution D_μ is c-bounded for $c \in (0,1]$. Let $\tilde{\mu} \in (-1,1)^n$ satisfy $\|\mu - \tilde{\mu}\|_\infty \leq \varepsilon$. Then the corresponding biased quantum Fourier transforms satisfy*

$$\left\| H_\mu - H_{\tilde{\mu}} \right\| \leq 2\sqrt{2} n \gamma \varepsilon,$$

where $\gamma = \frac{1}{c^2}\left((2-c)\frac{3}{2\sqrt{2}c} + 1\right)$.

Proof This proof is given in "Appendix B." □

The next Lemma is on approximating the bias parameter of an unknown product distribution from examples. (Compare the closing remark in Appendix A of [17].)

Lemma 10 *Using $m \leq \mathcal{O}(\frac{8\gamma^2 \cdot n^2}{\varepsilon^2} \ln(\frac{n}{\delta}))$ copies of the quantum example state $|\psi_a\rangle$ (or of $|\psi_a^{noisy}\rangle$) for a product distribution D_μ with bias vector $\mu \in (-1,1)^n$ s.t. D_μ is c-bounded for $c \in (0,1]$ one can, with probability $\geq 1-\delta$, output $\tilde{\mu} \in (-1,1)^n$ s.t. $\left\| H_\mu - H_{\tilde{\mu}} \right\| \leq \varepsilon$.*

Proof Recall that $\mu_i = \mathbb{E}_{D_\mu}[x_i]$. Via a standard application of Hoeffding's inequality we conclude that $\mathcal{O}(\frac{8\gamma^2 \cdot n^2}{\varepsilon^2} \ln(\frac{1}{\delta}))$ examples drawn i.i.d. from D_μ (which can be obtained from copies of the quantum example state by measuring the corresponding subsystem) are sufficient to guarantee that, with probability $\geq 1-\delta$, the empirical estimate $\hat{\mu}_i$ satisfies $|\mu_i - \hat{\mu}_i| \leq \frac{\varepsilon}{2\sqrt{2}\gamma \cdot n}$. As each component of a copy of the quantum example state can be measured separately, we see —using the union bound, that $\mathcal{O}(\frac{8\gamma^2 \cdot n^2}{\varepsilon^2} \ln(\frac{n}{\delta}))$ copies of the (possibly noisy) quantum example state suffice to guarantee that, with probability $\geq 1-\delta$, it holds that $\left\| \mu - \hat{\mu} \right\|_\infty \leq \frac{\varepsilon}{2\sqrt{2}\gamma \cdot n}$. Now we can apply the previous Corollary to finish the proof. □

If we now combine this result with Theorem 9, we obtain a sample complexity upper bound for our learning problem without assuming the underlying distribution to be known in advance.

Corollary 3 *Let $|\psi_a\rangle = \sum_{x \in \{-1,1\}^n} \sqrt{D_\mu(x)}|x, f^{(a)}(x)\rangle$, $a \in \{0,1\}^n$, $\mu \in (-1,1)^n$ s.t. D_μ is c-bounded for some $c \in (0,1]$ satisfying $c > 1 - \sqrt{\frac{1-2\varepsilon}{2n}}$. Then there exists a quantum algorithm which, given access to*

$$\mathcal{O}\left(\frac{8\gamma^2 \cdot n^2}{\varepsilon^2} \ln\left(\frac{n}{\delta}\right) + \max\left\{\frac{1}{(1-2\varepsilon)^2}, \frac{1}{1-2(n(1-c)^2+\varepsilon)^2}\right\} \ln\left(\frac{1}{\delta}\right)\right)$$

copies of the quantum example state $|\psi_a\rangle$, with probability $\geq 1-\delta$, outputs the string a, without prior knowledge of the underlying distribution D_μ.

Note, however, that the learning algorithm does need to obtain the c-boundedness parameter c as input in advance, but this (in general) does not fix the underlying distribution. Observe also that—since Lemma 10 remains valid for noisy quantum examples—, even though we do not explicitly formulate the result of this subsection for noisy quantum training data, such a generalization is possible by combining the strategies presented in this and the previous subsections.

B Proofs

Proof of Lemma 2 We directly compute the state produced by the algorithm before the measurement is performed:

$$\begin{aligned}(H_\mu \otimes H)|\psi_f\rangle &= \sum_{x\in\{-1,1\}^n} \sum_{j\in\{0,1\}^n} \frac{1}{\sqrt{2}} D_\mu(x)\phi_{\mu,j}(x)\left(|j,0\rangle + (-1)^{f(x)}|j,1\rangle\right) \\ &= \frac{1}{\sqrt{2}} \sum_{j\in\{0,1\}^n} \underbrace{\mathbb{E}_{D_\mu}[\phi_{\mu,j}]}_{=\delta_{j,0\ldots0}} |j,0\rangle + \underbrace{\mathbb{E}_{D_\mu}[g\phi_{\mu,j}]}_{=\hat{g}_\mu(j)} |j,1\rangle.\end{aligned}$$

Hence, the computational basis measurement from step 3 of Algorithm 1 on the last qubit returns 1 with probability $\frac{1}{2}$ and if that is the case, the computational basis measurement on the first n qubits will return j with probability $\left(\hat{g}_\mu(j)\right)^2$, as claimed. □

Proof of Lemma 6 The proof is by direct computation using the Fourier expansion:

$$\begin{aligned}(HGH^{-1})(a,b) &= \frac{1}{2^n} \sum_{c,d\in\{0,1\}^n} (-1)^{c\cdot a + d\cdot b} g(c+d) \\ &= \frac{1}{2^n} \sum_{c,d,j\in\{0,1\}^n} (-1)^{c\cdot a + d\cdot b + j\cdot(c+d)} \hat{g}(j) \\ &= \frac{1}{2^n} \sum_{j\in\{0,1\}^n} \hat{g}(j) \underbrace{\sum_{c\in\{0,1\}^n} (-1)^{c\cdot(a+j)}}_{=2^n\delta_{a,j}} \underbrace{\sum_{d\in\{0,1\}^n} (-1)^{d\cdot(b+j)}}_{=2^n\delta_{b,j}} \\ &= 2^n \hat{g}(a)\delta_{a,b}.\end{aligned}$$

Unitarity of H can be checked easily by exploiting the same identity as in the second to last line of the previous computation. □

Proof of Corollary 1 Using Lemma 6 we can directly compute the diagonal entries of the matrix root and obtain

$$\sqrt{G}(a,a) = \left(H^{-1} \cdot \operatorname{diag}\left(\left\{\sqrt{2^n \hat{g}(j)} \mid j \in \{0,1\}^n\right\}\right) \cdot H\right)(a,a)$$

$$= \frac{1}{2^n} \sum_{j,k\in\{0,1\}^n} (-1)^{c\cdot j + d\cdot k} \sqrt{2^n \hat{g}(j)}\delta_{j,k}$$
$$= \frac{1}{\sqrt{2^n}} \sum_{j\in\{0,1\}^n} \sqrt{\hat{g}(j)}$$

for every $a \in \{0,1\}^n$. □

Proof of Lemma 7 As in the proof of Lemma 4, due to the product structure of all the relevant objects (here our assumption on the form of the noise enters), it suffices to consider the case $n = 1$ in detail. In this case, we have $f^{(a)}(x) = a\tilde{x}$, $g^{(a)}(x) = (-1)^{a\tilde{x}+\xi_x}$ for $\tilde{x} = \frac{1-x}{2}$, $\phi_{\mu,0}(x) = 1$, and $\phi_{\mu,1}(x) = \frac{x-\mu}{\sqrt{1-\mu^2}}$. (We leave out unnecessary indices to improve readability.) We compute

$$\begin{aligned}\hat{g}_\mu^{(a)}(j) &= \mathbb{E}_{D_\mu}[(-1)^{a\tilde{x}+\xi_x}\phi_{\mu,j}(x)] \\ &= \frac{1+\mu}{2}\cdot(-1)^{\xi_1}\cdot\phi_{\mu,j}(1) + \frac{1-\mu}{2}\cdot(-1)^{a+\xi_{-1}}\cdot\phi_{\mu,j}(-1).\end{aligned}$$

By plugging in we now obtain

$$\begin{aligned}\hat{g}_\mu^{(0)}(0) &= \frac{1+\mu}{2}\cdot(-1)^{\xi_1}\cdot 1 + \frac{1-\mu}{2}\cdot(-1)^{\xi_{-1}}\cdot 1, \\ \hat{g}_\mu^{(0)}(1) &= \frac{1+\mu}{2}\cdot(-1)^{\xi_1}\cdot\frac{1-\mu}{\sqrt{1-\mu^2}} + \frac{1-\mu}{2}\cdot(-1)^{\xi_{-1}}\cdot\frac{-1-\mu}{\sqrt{1-\mu^2}}, \\ \hat{g}_\mu^{(1)}(0) &= \frac{1+\mu}{2}\cdot(-1)^{\xi_1}\cdot 1 + \frac{1-\mu}{2}\cdot(-1)^{1+\xi_{-1}}\cdot 1, \\ \hat{g}_\mu^{(1)}(1) &= \frac{1+\mu}{2}\cdot(-1)^{\xi_1}\cdot\frac{1-\mu}{\sqrt{1-\mu^2}} + \frac{1-\mu}{2}\cdot(-1)^{1+\xi_{-1}}\cdot\frac{-1-\mu}{\sqrt{1-\mu^2}}.\end{aligned}$$

So with probability $(\eta^1)^2 + (1-\eta^1)^2 = 1 - 2\eta^1(1-\eta^1)$, namely if $\xi_1 = \xi_{-1} = b \in \{0,1\}$, we obtain

$$\hat{g}_\mu^{(0)}(0) = (-1)^b, \quad \hat{g}_\mu^{(0)}(1) = 0, \quad \hat{g}_\mu^{(1)}(0) = (-1)^b\mu, \quad \hat{g}_\mu^{(1)}(1) = (-1)^b\sqrt{1-\mu^2},$$

and with probability $2\eta^1(1-\eta^1)$, namely if $\xi_1 = b \neq \xi_{-1}$, we obtain

$$\hat{g}_\mu^{(0)}(0) = (-1)^b\mu, \quad \hat{g}_\mu^{(0)}(1) = (-1)^b\sqrt{1-\mu^2}, \quad \hat{g}_\mu^{(1)}(0) = (-1)^b, \quad \hat{g}_\mu^{(1)}(1) = 0.$$

Therefore we obtain: With probability $1-2\eta^1(1-\eta^1)$ the μ-biased Fourier coefficients satisfy

$$\hat{g}_\mu^{(a)}(j) = \begin{cases}(-1)^b(1-j), & \text{for } a = 0 \\ (-1)^b((1-j)\mu + j\sqrt{1-\mu^2}) & \text{for } a = 1\end{cases},$$

and with probability $2\eta^1(1-\eta^1)$ the μ-biased Fourier coefficients satisfy

$$\hat{g}_\mu^{(a)}(j) = \begin{cases} (-1)^b((1-j)\mu + j\sqrt{1-\mu^2}) & \text{for } a = 0 \\ (-1)^b(1-j), & \text{for } a = 1 \end{cases},$$

which is exactly the claim for $n = 1$. □

Proof of Theorem 8 We want to prove that $\mathbb{P}[\text{Algorithm 6 does not output } a] \leq \delta$, where the probability is w.r.t. both the internal randomness of the algorithm and the random variables.

First observe that, due to (i) in Theorem 7, exactly the same reasoning as in the proof of Theorem 4 shows that the probability of observing $j_{n+1} = 1$ in at most $k-1$ of the m runs of Algorithm 5 (assuming $k \leq \frac{m}{2}$) is bounded by

$$\mathbb{P}\left[\text{Bin}\left(m, \frac{1}{2}\right) \geq m - k\right] \leq \exp\left(-\frac{2\left(\frac{m}{2} - k\right)^2}{m}\right). \tag{B.1}$$

We will now search for the number of observations of $j_{n+1} = 1$ which is required to guarantee that the majority string is correct with high probability. Suppose we observe $j_{n+1} = 1$ in k runs of Algorithm 5, $k \in 2\mathbb{N}$. Again we see that

$$\begin{aligned}\mathbb{P}[\exists 1 \leq i \leq n : a_i \neq o_i] \leq\ & \mathbb{P}[\exists 1 \leq i \leq n : a_i = 0 \neq o_i] \\ & + \mathbb{P}[\exists 1 \leq i \leq n : a_i = 1 \neq o_i].\end{aligned}$$

As "false 1's" can only appear in the case where our noise variables have an influence (compare Theorem 7), we will first find a lower bound on k which guarantees that the probability of the noise variable influence becoming relevant for at least $\frac{k}{5}$ runs is $\leq \frac{\delta}{4}$. Namely, we bound (again via Hoeffding)

$$\begin{aligned}\mathbb{P}\left[\text{Bin}(k, n\rho) \geq \frac{k}{5}\right] &= \mathbb{P}\left[\text{Bin}(k, n\rho) - kn\rho \geq k\left(\frac{1}{5} - n\rho\right)\right] \\ &\leq \exp\left(-2k\left(\frac{1-5n\rho}{5}\right)^2\right).\end{aligned}$$

We now set this last expression $\leq \frac{\delta}{4}$ and rearrange the inequality to

$$k \geq \frac{25}{2(1-5n\rho)^2} \ln\left(\frac{4}{\delta}\right).$$

Now we will find a lower bound on k which guarantees that, if the noise variable influence is relevant in at most $\frac{k}{5}$ of the runs, among the remaining $\frac{4k}{5}$ runs with probability $\geq 1 - \frac{\delta}{4}$ we make at most $\frac{k}{5}$ "false 0" observations. To this end, we bound

(again via Hoeffding)

$$\begin{aligned}
&\mathbb{P}\left[\mathrm{Bin}\left(\frac{4k}{5}, n(1-c)^2\right) \geq \frac{k}{5}\right] \\
&\quad = \mathbb{P}\left[\mathrm{Bin}\left(\frac{4k}{5}, n(1-c)^2\right) - \frac{4kn(1-c)^2}{5} \geq \frac{k}{5} - \frac{4kn(1-c)^2}{5}\right] \\
&\quad \leq \exp\left(-2k\left(\frac{1}{5} - \frac{4n(1-c)^2}{5}\right)^2\right).
\end{aligned}$$

We now set this last expression $\leq \frac{\delta}{4}$ and rearrange the inequality to

$$k \geq \frac{25}{2(1-4n(1-c)^2)^2} \ln\left(\frac{4}{\delta}\right).$$

Hence, by the union bound a sufficient condition for $\mathbb{P}[\exists 1 \leq i \leq n : a_i \neq o_i] \leq \frac{\delta}{2}$ to hold is given by

$$k \geq \frac{25}{2} \max\left\{\frac{1}{(1-5n\rho)^2}, \frac{1}{(1-4n(1-c)^2)^2}\right\} \ln\left(\frac{4}{\delta}\right). \tag{B.2}$$

Combining Eqs. (B.2) and (B.1) we now require

$$\exp\left(-\frac{2\left(\frac{25}{2}\max\left\{\frac{1}{(1-5n\rho)^2}, \frac{1}{(1-4n(1-c)^2)^2}\right\}\ln\left(\frac{4}{\delta}\right) - \frac{m}{2}\right)^2}{m}\right) \overset{!}{\leq} \frac{\delta}{4}.$$

Rearranging gives the sufficient condition

$$m \geq 25 \max\left\{\frac{1}{(1-5n\rho)^2}, \frac{1}{(1-4n(1-c)^2)^2}\right\} \ln\left(\frac{4}{\delta}\right).$$

This proves the claim of the theorem thanks to the union bound. □

Proof of Corollary 2 According to the Lemma 9 it holds that

$$\begin{aligned}
&\left\|H_\mu - H_{\tilde{\mu}}\right\| \\
&\quad \leq \sum_{i=1}^{n} \left\|\mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes H_{\mu_i} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1} - \mathbb{1} \otimes \ldots \otimes \mathbb{1} \otimes H_{\tilde{\mu}_i} \otimes \mathbb{1} \otimes \ldots \otimes \mathbb{1}\right\| \\
&\quad = \sum_{i=1}^{n} \left\|H_{\mu_i} - H_{\tilde{\mu}_i}\right\|.
\end{aligned}$$

Thus it suffices to bound the operator norm of the difference of the 1-qubit biased quantum Fourier transforms. So let $|\varphi\rangle = \sum_{x\in\{-1,1\}} \alpha_x |x\rangle$ be a qubit state. Then

$$(H_{\mu_j} - H_{\tilde{\mu}_j})|\varphi\rangle = \sum_{x\in\{-1,1\}} \sum_{j\in\{0,1\}} \left(\sqrt{D_{\mu_i}(x)}\phi_{\mu_i,j}(x) - \sqrt{D_{\tilde{\mu}_i}(x)}\phi_{\tilde{\mu}_i,j}(x)\right)\alpha_x|j\rangle.$$

We have to bound the (Euclidean) norm of this vector. To achieve this, we will bound (for arbitrary $x \in \{-1, 1\}$ and $j \in \{0, 1\}$) the expression

$$\left|\sqrt{D_{\mu_i}(x)}\phi_{\mu_i,j}(x) - \sqrt{D_{\tilde{\mu}_i}(x)}\phi_{\tilde{\mu}_i,j}(x)\right|^2.$$

This is done by direct computation using $1 - \mu_i^2 \geq 1 - (1-c)^2 \geq c^2$, $1 - \tilde{\mu}_i^2 \geq c^2$ and $|\mu_i - \tilde{\mu}_i| \leq \varepsilon$ as follows:

$$\begin{aligned}
&\left|\sqrt{D_{\mu_i}(x)}\phi_{\mu_i,j}(x) - \sqrt{D_{\tilde{\mu}_i}(x)}\phi_{\tilde{\mu}_i,j}(x)\right| \\
&= \left|\frac{(x_i-\mu_i)\sqrt{1-\tilde{\mu}_i^2}\sqrt{D_{\mu_i}(x)} - (x_i-\tilde{\mu}_i)\sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}}{\sqrt{1-\tilde{\mu}_i^2}\sqrt{1-\mu_i^2}}\right| \\
&\leq \frac{1}{c^2}\left|(x_i-\mu_i)\sqrt{1-\tilde{\mu}_i^2}\sqrt{D_{\mu_i}(x)} - (x_i-\tilde{\mu}_i)\sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}\right| \\
&= \frac{1}{c^2}\left|(x_i-\mu_i)\left(\sqrt{1-\tilde{\mu}_i^2}\sqrt{D_{\mu_i}(x)} - \sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}\right)\right. \\
&\quad \left. + (\tilde{\mu}_i-\mu_i)\sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}\right| \\
&\leq \frac{1}{c^2}\left(\left|(x_i-\mu_i)\left(\sqrt{1-\tilde{\mu}_i^2}\sqrt{D_{\mu_i}(x)} - \sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}\right)\right|\right. \\
&\quad \left. + \left|(\tilde{\mu}_i-\mu_i)\sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}\right|\right) \\
&\leq \frac{1}{c^2}\left((2-c)\left|\sqrt{1-\tilde{\mu}_i^2}\sqrt{D_{\mu_i}(x)} - \sqrt{1-\mu_i^2}\sqrt{D_{\tilde{\mu}_i}(x)}\right| + \varepsilon\right) \\
&\leq \frac{1}{c^2}\left((2-c)\left(\left|\sqrt{D_{\mu_i}(x)} - \sqrt{D_{\tilde{\mu}_i}(x)}\right| + \left|\sqrt{1-\mu_i^2} - \sqrt{1-\tilde{\mu}_i^2}\right|\right) + \varepsilon\right).
\end{aligned}$$

Now note that

$$\begin{aligned}
\left|\left(\sqrt{D_{\mu_i}(x)} - \sqrt{D_{\tilde{\mu}_i}(x)}\right)\left(\sqrt{D_{\mu_i}(x)} + \sqrt{D_{\tilde{\mu}_i}(x)}\right)\right| &= \left|D_{\mu_i}(x) - D_{\tilde{\mu}_i}(x)\right| \\
&= \left|\frac{1+\tilde{x}_i\mu_i}{2} - \frac{1+\tilde{x}_i\tilde{\mu}_i}{2}\right| \\
&= \frac{1}{2}|\mu_i - \tilde{\mu}_i|,
\end{aligned}$$

which implies

$$\begin{aligned}\left|\sqrt{D_{\mu_i}(x)}-\sqrt{D_{\tilde{\mu}_i}(x)}\right| &= \left|\frac{\mu_i-\tilde{\mu}_i}{2\left(\sqrt{D_{\mu_i}(x)}+\sqrt{D_{\tilde{\mu}_i}(x)}\right)}\right| \\ &\leq \frac{\varepsilon}{2}\frac{1}{2\sqrt{\frac{c}{2}}} \\ &= \frac{\varepsilon}{2\sqrt{2c}},\end{aligned}$$

and that moreover

$$\begin{aligned}\left|\left(\sqrt{1-\mu_i^2}-\sqrt{1-\tilde{\mu}_i^2}\right)\left(\sqrt{1-\mu_i^2}+\sqrt{1-\tilde{\mu}_i^2}\right)\right| &= \left|1-\mu_i^2-(1-\tilde{\mu}_i^2)\right| \\ &= \left|\mu_i^2-\tilde{\mu}_i^2\right|,\end{aligned}$$

which in turn implies

$$\begin{aligned}\left|\sqrt{1-\mu_i^2}-\sqrt{1-\tilde{\mu}_i^2}\right| &= \left|\frac{\mu_i^2-\tilde{\mu}_i^2}{\sqrt{1-\mu_i^2}+\sqrt{1-\tilde{\mu}_i^2}}\right| \\ &\leq \frac{|\mu_i+\tilde{\mu}_i|\cdot|\mu_i-\tilde{\mu}_i|}{2\sqrt{1-(1-c)^2}} \\ &\leq \frac{2\varepsilon}{2\sqrt{2c-c^2}} \\ &\leq \frac{\varepsilon}{\sqrt{2}c}.\end{aligned}$$

Hence, we obtain

$$\left|\sqrt{D_{\mu_i}(x)}\phi_{\mu_i,j}(x)-\sqrt{D_{\tilde{\mu}_i}(x)}\phi_{\tilde{\mu}_i,j}(x)\right| \leq \frac{1}{c^2}\left((2-c)\left(\frac{\varepsilon}{2\sqrt{2c}}+\frac{\varepsilon}{\sqrt{2}c}\right)+\varepsilon\right) \leq \gamma\varepsilon,$$

where we defined $\gamma := \frac{1}{c^2}\left((2-c)\frac{3}{2\sqrt{2}c}+1\right)$. This now implies

$$\begin{aligned}\left\|(H_{\mu_j}-H_{\tilde{\mu}_j})|\varphi\rangle\right\|_2 &\leq \sum_{x\in\{-1,1\}}\sum_{j\in\{0,1\}}\left\|\left(\sqrt{D_{\mu_i}(x)}\phi_{\mu_i,j}(x)-\sqrt{D_{\tilde{\mu}_i}(x)}\phi_{\tilde{\mu}_i,j}(x)\right)\alpha_x|j\rangle\right\|_2 \\ &\leq \gamma\varepsilon\sum_{x\in\{-1,1\}}\sum_{j\in\{0,1\}}|\alpha_x| \\ &= 2\gamma\varepsilon\sum_{x\in\{-1,1\}}|\alpha_x| \\ &\leq 2\sqrt{2}\gamma\varepsilon.\end{aligned}$$

Finally, we get

$$\left\| H_\mu - H_{\tilde{\mu}} \right\| \leq \sum_{i=1}^{n} \left\| H_{\mu_i} - H_{\tilde{\mu}_i} \right\| \leq 2\sqrt{2} n \gamma \varepsilon,$$

as claimed. □

References

1. Arunachalam, S., Chakraborty, S., Lee, T., Paraashar, M., de Wolf, R.: Two new results about quantum exact learning (2018). arXiv:1810.00481
2. Arunachalam, S., Grilo, A.B., Sundaram, A.: Quantum hardness of learning shallow classical circuits (2019). arXiv:1903.02840
3. Arunachalam, S., de Wolf, R.: Guest column: a survey of quantum learning theory. SIGACT News **48**, (2017). https://doi.org/10.1145/3106700.3106710. https://pure.uva.nl/ws/files/25255496/p41_arunachalam.pdf
4. Arunachalam, S., de Wolf, R.: Optimal quantum sample complexity of learning algorithms. J. Mach. Learn. Res. **19**(71), 1–36 (2018). http://jmlr.org/papers/v19/18-195.html
5. Atıcı, A., Servedio, R.A.: Quantum algorithms for learning and testing juntas. Quantum Inf. Process. **6**(5), 323–348 (2007). https://doi.org/10.1007/s11128-007-0061-6
6. Barnum, H., Knill, E.: Reversing quantum dynamics with near-optimal quantum and classical fidelity (2000). arXiv:quant-ph/0004088
7. Benioff, P.: The computer as a physical system: a microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. J. Stat. Phys. **22**(5), 563–591 (1980). https://doi.org/10.1007/BF01011339
8. Bernstein, E., Vazirani, U.: Quantum complexity theory. In: Kosaraju R. (ed.) Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, pp. 11–20. ACM, New York (1993). https://doi.org/10.1145/167088.167097
9. Blum, A., Kalai, A., Wasserman, H.: Noise-tolerant learning, the parity problem, and the statistical query model. J. ACM **50**(4), 506–519 (2003). https://doi.org/10.1145/792538.792543
10. Bshouty, N.H., Jackson, J.C.: Learning dnf over the uniform distribution using a quantum example oracle. SIAM J. Comput. **28**(3), 1136–1153 (1998). https://doi.org/10.1137/S0097539795293123
11. Cross, A.W., Smith, G., Smolin, J.A.: Quantum learning robust against noise. Phys. Rev. A **92**(1), 97 (2015). https://doi.org/10.1103/PhysRevA.92.012327
12. Feynman, R.P.: Quantum mechanical computers. Opt. News **11**(2), 11 (1985). https://doi.org/10.1364/ON.11.2.000011
13. Grilo, A.B., Kerenidis, I., Zijlstra, T.: Learning with errors is easy with quantum samples (2017). Phys. Rev. A **99**(3), 032314 (2019). https://doi.org/10.1103/PhysRevA.99.032314
14. Hausladen, P., Wootters, W.K.: A 'pretty good' measurement for distinguishing quantum states. J. Mod. Opt. **41**(12), 2385–2390 (1994). https://doi.org/10.1080/09500349414552221
15. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830
16. Ivanyos, G., Prakash, A., Santha, M. (eds.): On Learning Linear Functions from Subset and Its Applications in Quantum Computing. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern (2018). https://doi.org/10.4230/LIPICS.ESA.2018.66
17. Kanade, V., Rocchetto, A., Severini, S.: Learning dnfs under product distributions via μ-biased quantum fourier sampling. Quantum Inf. Comput. **19**(15&16), 1261–1278 (2019)
18. Lyubashevsky, V.: The parity problem in the presence of noise, decoding random linear codes, and the subset sum problem. In: Chekuri C. (ed.) Approximation, Randomization and Combinatorial optimization. Lecture Notes in Computer Science, vol. 3624, pp. 378–389. Springer, Berlin (2005). https://doi.org/10.1007/11538462_32
19. Montanaro, A.: The quantum query complexity of learning multilinear polynomials. Inf. Process. Lett. **112**(11), 438–442 (2012). https://doi.org/10.1016/j.ipl.2012.03.002

20. Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information, 10 Anniversary edn. Cambridge University Press, Cambridge (2010)
21. O'Donnell, R.: Analysis of Boolean Functions. Cambridge University Press, Cambridge (2014)
22. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. J. ACM **56**(6), 1–40 (2009). https://doi.org/10.1145/1568318.1568324
23. Ristè, D., da Silva, M.P., Ryan, C.A., Cross, A.W., Córcoles, A.D., Smolin, J.A., Gambetta, J.M., Chow, J.M., Johnson, B.R.: Demonstration of quantum advantage in machine learning. npj Quantum Inf. **3**(1), 16 (2017). https://doi.org/10.1038/s41534-017-0017-3
24. Servedio, R.A., Gortler, S.J.: Equivalences and separations between quantum and classical learnability. SIAM J. Comput. **33**(5), 1067–1092 (2004). https://doi.org/10.1137/S0097539704412910
25. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: from theory to algorithms. Cambridge University Press, Cambridge (2014)
26. Valiant, L.G.: A theory of the learnable. Commun. ACM **27**(11), 1134–1142 (1984). https://doi.org/10.1145/1968.1972
27. Vershynin, R.: High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 47. Cambridge University Press, Cambridge (2018)

Affiliations

Matthias C. Caro[1]

[1] Department of Mathematics, Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany

B.2 Generalization in quantum machine learning from few training data

Generalization in quantum machine learning from few training data

Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles

---

Whether variational quantum machine learning models (QMLMs) can serve as a relevant area of application for quantum computing in the near term depends, among other things, crucially on their training data requirements. In particular, it is important to understand how many training data points suffice to guarantee good generalization for QMLMs. This work proves broadly applicable upper bounds on the training data size sufficient for a QMLM to generalize from training data to new data. Moreover, we showcase our theoretical guarantees for two applications of variational QML and confirm them in numerical experiments.

After an introductory Section I, we present the main results of our paper in Section II. Subsection II.A establishes our mathematical framework: We consider QMLMs as described by parametrized CPTP maps $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}$, where we allow for continuous parameters in trainable gates and for discrete parameters for variable circuit structure. With such a QMLM, we act on a subsystem of an input state and then measure a data-dependent loss observable to obtain a loss function given as $\ell(\boldsymbol{\alpha}; x_i, y_i) = \mathrm{tr}\left[O_{x_i,y_i}^{\mathrm{loss}}(\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id})(\rho(x_i))\right]$. In the spirit of probably approximately correct (PAC) learning, the goal is now to achieve a small expected risk with respect to this loss and the (unknown) data-generating probability measure. In Subsection II.B, we give informal statements of our theoretical results, which are PAC generalization bounds for QMLMs, and discuss some of their implications. This is followed by our numerical results in Section II.C. Here, Subsection II.C.1 deals with quantum phase recognition using a specific QMLM, namely a quantum convolutional neural network (QCNN). And in Subsection II.C.2, we employ a variable-structure QMLM for unitary compiling. Both of these numerical investigations confirm our theoretical generalization bounds and suggest an even more favorable generalization behavior for certain cases.

We further discuss our results and findings in Section III. First, in Subsection III.A, we describe potential further areas of application for our generalization guarantees. Next, we compare our results to prior work in Subsection III.B. And in Subsection III.C, we elaborate the potential relevance of our results to the quest for quantum advantage in machine learning and present some open questions. Finally, Section IV gives an overview over our methods, where Subsection III.A focuses on the theoretical results and Subsection III.B explains the numerical experiments in more detail.

Appendix A contains a detailed review of related work, which subdivide into discussions of prior work on generalization bounds for variational QML, on quantum phase recognition, and on unitary compiling. In Appendix B, we collect some auxiliary Lemmata from statistical learning theory and from quantum information theory that are important tools for proving our results.

Appendix C contains all the mathematical results and proofs of the paper. In Section C.1, we prove covering number bounds for parametrized quantum circuits in terms of the number of trainable local gates. More precisely, in Theorems 6 and 7, combining a basic covering number bound for 2-qubit quantum gates and the subadditivity of the distance induced by the diamond

norm, we bound the ε-metric entropy of the class of n-qubit CPTP maps that a QMLM with T trainable local quantum gates can implement by $\mathcal{O}(T\log(T/\varepsilon))$, where distance is measured using the diamond norm. Crucially, this upper bound is independent of n, the number of qubits, and scales only slightly superlinearly in T, the number of trainable gates. Theorem 8 extends our covering number bounds to QMLM architectures in which the same parametrized gates are reused multiple times.

We present our main mathematical results in detail in Section C.2. Theorem 10 in Section C.2.1 provides a simple proof for a generalization bound derived from our covering number bounds, combining Hoeffding's inequality with a union bound over elements in a covering net. The obtained generalization bound, however, has a suboptimal dependence on the training data size N. In Subsection C.2.2, we show how to improve the N-dependence using Dudley's Theorem and a generalization bound in terms of Rademacher complexities. This allows us to prove Theorem 11: With high probability, the generalization error of a QMLM with T trainable local gates behaves as $\mathcal{O}(\sqrt{T\log(T)/N})$, when training on data of size N. The next subsections deal with different extensions of this result. First, in Subsection C.2.3, we allow for QMLMs in which the same trainable gates are used multiple times. As we show in Corollaries 1 and 2, the generalization error scales at worst logarithmically with the number of uses per trainable gate. Next, Subsection C.2.4 considers QMLMs with variable circuit architecture, so that also the number and placement of trainable gates can be optimized during training. Corollary 3 establishes that the generalization performance depends at worst logarithmically on the number of different architectures considered during the optimization. Theorem 12 in Subsection C.2.5 provides an optimization-dependent tightening of the generalization guarantee of Theorem 11, assuming that some of the trainable gates do not change much during training. The proof of Theorem 12 is based on an extension of the covering number bounds of Section C.1, which we give in Theorem 13. As our final extension, we describe in Corollary 4 of Subsection C.2.6 how using an unbiased estimator for evaluating the training error influences the generalization. Finally, in Subsection C.2.7, we summarize all of these extensions in Theorem 9, the most general form of our results.

Appendices D and E contain a detailed explanation of how our theory applies to the two applications that we also investigate numerically. First, in Appendix D, we show how to phrase the problem of quantum phase recognition in our framework. And we demonstrate that for QCNNs, our results guarantee good generalization already from training data size growing only polylogarithmically with the system size. Second, Appendix E describes a variational QML approach to unitary compiling. Here, our results imply that polynomial-size training data suffices for good generalization, assuming that the target unitary to be compiled can be implemented using polynomially many local gates.

I was significantly involved in finding the ideas and carrying out the scientific work of all parts of this article, with the exception of the numerical experiments. The idea for this project goes back to a suggestion by Hsin-Yuan Huang and was further developed in discussions between Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, Patrick J. Coles, and myself. I was significantly involved in writing the main part of the paper, with the exception of Section IV.B. I wrote the majority of the technical Appendix of the paper, with the exception of Subsections A.2 and A.3.

Permission to include:

Matthias C. Caro, Hsin-Yuan Huang, M. Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles.
Generalization in quantum machine learning from few training data.
Nature Communications 13, 4919 (2022). https://doi.org/10.1038/s41467-022-32550-3.

Reprints & Permissions
Author reprints
Commercial reprints
Permissions requests
Other services
Frequently asked questions
Contact details

Permissions requests

Springer Nature grants permission for authors, readers and third parties to reproduce material from its journals and online products as part of another publication or entity. This includes, for example, the use of a figure in a presentation, the posting of an abstract on a web site, or the reproduction of a full article within another journal. Certain permissions can be granted free of charge; others incur a fee.

On this page

- Types of permission request
- Get permission to reuse Springer Nature content online
  - Permission requests from authors
  - Self-archiving
  - Author reuse
  - Get permission to reuse Springer Nature content online
  - How to obtain permission to reuse Springer Nature content not available online

For answers to frequently asked questions click here.

Types of permission request

Permission can be obtained for re-use of portions of material - ranging from a single figure to a whole paper - in books, journals/magazines, newsletters, theses/dissertations, classroom materials/academic course packs, academic conference materials, training materials (including continuing medical education),

promotional materials, and web sites. Some permission requests can be granted free of charge, others carry a fee.

Springer Nature does not allow PDFs of full papers to be reproduced online, however e-print PDFs can be purchased as commercial reprints. If you wish to purchase multiple stand-alone copies of a Nature Portfolio paper, which is then printed and shipped to you, please go to commercial reprints.

Top of page ⤴

Get permission to reuse Springer Nature content online

Permission requests from authors

The author of articles published by Springer Nature do not usually need to seek permission for re-use of their material as long as the journal is credited with initial publication.

Ownership of copyright in in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:

To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).

The author and any academic institution where they work at the time may reproduce the contribution for the purpose of course teaching.

To reuse figures or tables created by the Author and contained in the Contribution in oral presentations and other works created by them.

To post a copy of the contribution as accepted for publication after peer review (in locked Word processing file, of a PDF version thereof) on the Author's own web site, or the Author's institutional repository, or the Author's funding body's archive, six months after publication of the printed or online edition of the Journal, provided that they also link to the contribution on the publisher's website.

The above use of the term 'Contribution' refers to the author's own version, not the final version as published in the Journal.

Self-archiving

Authors retain the right to self-archive the final accepted version of their manuscript. Please see our self-archiving policy for full details: http://www.nature.com/authors/policies/license.html

Author reuse

Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution.

Authors must properly cite the published article in their thesis according to current citation standards.

Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]'

If you are any doubt about whether your intended re-use is covered, please contact journalpermissions@springernature.com for confirmation.

Get permission to reuse Springer Nature content online

Springer Nature is partnered with the Copyright Clearance Center to meet our customers' licensing and permissions needs.

Copyright Clearance Center's RightsLink® service makes it faster and easier to secure permission for the reuse of Springer Nature content.

- Simply visit SpringerLink or www.nature.com and locate the desired content;
- Once you have opened the article or book chapter click on the "Rights and Permissions" button. This can either be found under the article title, at the bottom of the page or in the tools menu.
- Select the way you would like to reuse the content;
- Create an account if you haven't already;
- Accept the terms and conditions and you're done!

For questions about using the RightsLink service, please contact Customer Support at Copyright Clearance Center via phone +1-855-239-3415 or +1-978-646-2777 or email customercare@copyright.com.

How to obtain permission to reuse Springer Nature content not available online

Requests for permission to reuse content (e.g. figure or table, abstract, text excerpts) from Springer Nature publications currently

not available online must be submitted in writing to the following email addresses:

For English language Journal Permission queries please contact journalpermissions@springernature.com

For Books and German language Journal Permission queries please contact bookpermissions@springernature.com

Top of page ⤴

Nature Portfolio

nature communications

Article

https://doi.org/10.1038/s41467-022-32550-3

Generalization in quantum machine learning from few training data

Received: 12 April 2022

Accepted: 4 August 2022

Published online: 22 August 2022

Check for updates

Matthias C. Caro[1,2] ✉, Hsin-Yuan Huang[3,4], M. Cerezo[5,6], Kunal Sharma[7], Andrew Sornborger[5,8], Lukasz Cincio[9] & Patrick J. Coles[9]

Modern quantum machine learning (QML) methods involve variationally optimizing a parameterized quantum circuit on a training data set, and subsequently making predictions on a testing data set (i.e., generalizing). In this work, we provide a comprehensive study of generalization performance in QML after training on a limited number N of training data points. We show that the generalization error of a quantum machine learning model with T trainable gates scales at worst as $\sqrt{T/N}$. When only $K \ll T$ gates have undergone substantial change in the optimization process, we prove that the generalization error improves to $\sqrt{K/N}$. Our results imply that the compiling of unitaries into a polynomial number of native gates, a crucial application for the quantum computing industry that typically uses exponential-size training data, can be sped up significantly. We also show that classification of quantum states across a phase transition with a quantum convolutional neural network requires only a very small training data set. Other potential applications include learning quantum error correcting codes or quantum dynamical simulation. Our work injects new hope into the field of QML, as good generalization is guaranteed from few training data.

The ultimate goal of machine learning (ML) is to make accurate predictions on unseen data. This is known as generalization, and significant effort has been expended to understand the generalization capabilities of classical ML models. For example, theoretical results have been formulated as upper bounds on the generalization error as a function of the training data size and the model complexity[1–5]. Such bounds provide guidance as to how much training data is required and/or sufficient to achieve accurate generalization.

Quantum machine learning (QML) is an emerging field that has generated great excitement[6–9]. Modern QML typically involves training a parameterized quantum circuit in order to analyze either classical or quantum data sets[10–16]. Early results indicate that, for classical data analysis, QML models may offer some advantage over classical models under certain circumstances[17–19]. It has also been proven that QML models can provide an exponential advantage in sample complexity for analyzing quantum data[20,21].

However, little is known about the conditions needed for accurate generalization in QML. Significant progress has been made in understanding the trainability of QML models[18,22–36], but trainability is a separate question from generalization[18,37,38]. Overfitting of training data could be an issue for QML, just as it is for classical machine learning. Moreover, the training data size required for QML generalization has yet to be fully studied. Naïvely, one could expect that an exponential number of training points are needed when training a function acting on an exponentially large Hilbert space. For instance, some studies have found that, exponentially in n, the number of

[1]Department of Mathematics, Technical University of Munich, Garching, Germany. [2]Munich Center for Quantum Science and Technology (MCQST), Munich, Germany. [3]Institute for Quantum Information and Matter, Caltech, Pasadena, CA, USA. [4]Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA, USA. [5]Information Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [6]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [7]Joint Center for Quantum Information and Computer Science, University of Maryland, College Park, MD 20742, USA. [8]Quantum Science Center, Oak Ridge, TN 37931, USA. [9]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. ✉e-mail: caro@ma.tum.de

qubits, large amounts of training data would be needed, assuming that one is trying to train an arbitrary unitary[39,40]. This is a concerning result, since it would imply exponential scaling of the resources required for QML, which is precisely what the field of quantum computation would like to avoid.

In practice, a more relevant scenario to consider instead of arbitrary unitaries is learning a unitary that can be represented by a polynomial-depth quantum circuit. This class of unitaries corresponds to those that can be efficiently implemented on a quantum computer, and it is exponentially smaller than that of arbitrary unitaries. More generally, one could consider a QML model with T parameterized gates and relate the training data size N needed for generalization to T. Even more general would be to consider generalization error a dynamic quantity that varies during the optimization.

In this work, we prove highly general theoretical bounds on the generalization error in variational QML: The generalization error is approximately upper bounded by $\sqrt{T/N}$. In our proofs, we first establish covering number bounds for the class of quantum operations that a variational QML model can implement. From these, we then derive generalization error bounds using the chaining technique for random processes. A key implication of our results is that an efficiently implementable QML model, one such that $T \in \mathcal{O}(\mathrm{poly}n)$, only requires an efficient amount of training data, $N \in \mathcal{O}(\mathrm{poly}n)$, to obtain good generalization. This implication, by itself, will improve the efficiency guarantees of variational quantum algorithms[10,41,42] that employ training data, such as quantum autoencoders[13], quantum generative adversarial networks[43], variational quantum error correction[44,45], variational quantum compiling[46,47], and variational dynamical simulation[48–51]. It also yields improved efficiency guarantees for classical algorithms that simulate QML models.

We furthermore refine our bounds to account for the optimization process. We show that generalization improves if only some parameters have undergone substantial change during the optimization. Hence, even if we used a number of parameters T larger than the training data size N, the QML model could still generalize well if only some of the parameters have changed significantly. This suggests that QML researchers should be careful not to overtrain their models especially when the decrease in training error is insufficient.

To showcase our results, we consider quantum convolutional neural networks (QCNNs)[27,45], a QML model that has received significant attention. QCNNs have only $T = \mathcal{O}(\log n)$ parameters and yet they are capable of classifying quantum states into distinct phases. Our theory guarantees that QCNNs have good generalization error for quantum phase recognition with only polylogarithmic training resources, $N \in \mathcal{O}(\log^2 n)$. We support this guarantee with a numerical demonstration, which suggests that even constant-size training data can suffice.

Finally, we highlight the task of quantum compiling, a crucial application for the quantum computing industry. State-of-the-art classical methods for approximate optimal compiling of unitaries often employ exponentially large training data sets[52–54]. However, our work indicates that only polynomial-sized data sets are needed, suggesting that state-of-the-art compilers could be further improved. Indeed, we numerically demonstrate the surprisingly low data cost of compiling the quantum Fourier transform at relatively large scales.

Results

Framework

Let us first outline our theoretical framework. We consider a quantum machine learning model (QMLM) as being a parameterized quantum channel, i.e., a completely positive trace preserving (CPTP) map that is parameterized. We denote a QMLM as $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ where $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{k})$ denotes the set of parameters, including continuous parameters $\boldsymbol{\theta}$ inside gates, as well as discrete parameters $\boldsymbol{k}$ that allow the gate structure to vary. We make no further assumptions on the form of the dependence of the CPTP map $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ on the parameters $\boldsymbol{\alpha}$. During the training process, one would optimize the continuous parameters $\boldsymbol{\theta}$ and potentially also the structure $\boldsymbol{k}$ of the QMLM.

A QMLM takes input data in the form of quantum states. For classical data x, the input is first encoded in a quantum state via a map $x \mapsto \rho(x)$. This allows the data to be either classical or quantum in nature, since regardless it is eventually encoded in a quantum state. We assume that the data encoding is fixed in advance and not optimized over. We remark here that our results also apply for more general encoding strategies involving data re-uploading[55], as we explain in Supplementary Note 3.

For the sake of generality, we allow the QMLM to act on a subsystem of the state $\rho(x)$. Hence, the output state can be written as $(\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id})(\rho(x))$. For a given data point (x_i, y_i), we can write the loss function as

$$\ell(\boldsymbol{\alpha}; x_i, y_i) = \mathrm{Tr}\left[O_{x_i,y_i}^{\mathrm{loss}}\left(\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id}\right)(\rho(x_i))\right], \tag{1}$$

for some Hermitian observable $O_{x_i,y_i}^{\mathrm{loss}}$. As is common in classical learning theory, the prediction error bounds will depend on the largest (absolute) value that the loss function can attain. In our case, we therefore assume $C_{\mathrm{loss}} := \sup_{x,y} ||O_{x,y}^{\mathrm{loss}}|| < \infty$<?A3B2tal?>, i.e., the spectral norm can be bounded uniformly over all possible loss observables.

In Eq. (1), we take the measurement to act on a single copy of the output of the QMLM $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ upon input of (a subsystem of) the data encoding state $\rho(x_i)$. At first this looks like a restriction. However, note that one can choose $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ to be a tensor product of multiple copies of a QMLM, each with the same parameter setting, applied to multiple copies of the input state. Hence our framework is general enough to allow for global measurements on multiple copies. In this addition to the aforementioned situation, we further study the case in which trainable gates are more generally reused.

For a training dataset $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, the average loss for parameters $\boldsymbol{\alpha}$ on the training data is

$$\hat{R}_S(\boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{\alpha}; x_i, y_i), \tag{2}$$

which is often referred to as the *training error*. When we obtain a new input x, the *prediction error* of a parameter setting $\boldsymbol{\alpha}$ is taken to be the expected loss

$$R(\boldsymbol{\alpha}) = \mathop{\mathbb{E}}_{(x,y)\sim P}[\ell(\boldsymbol{\alpha}; x, y)], \tag{3}$$

where the expectation is with respect to the distribution P from which the training examples are generated.

Achieving small prediction error $R(\boldsymbol{\alpha})$ is the ultimate goal of (quantum) machine learning. As P is generally not known, the training error $\hat{R}_S(\boldsymbol{\alpha})$ is often taken as a proxy for $R(\boldsymbol{\alpha})$. This strategy can be justified via bounds on the *generalization error*

$$\mathrm{gen}(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha}) - \hat{R}_S(\boldsymbol{\alpha}), \tag{4}$$

which is the key quantity that we bound in our theorems.

Analytical results

We prove probabilistic bounds on the generalization error of a QMLM. Our bounds guarantee that a good performance on a sufficiently large training data set implies, with high probability, a good performance on previously unseen data points. In particular, we provide a precise meaning of "sufficiently large" in terms of properties of the QMLM and the employed training procedure.

https://doi.org/10.1038/s41467-022-32550-3

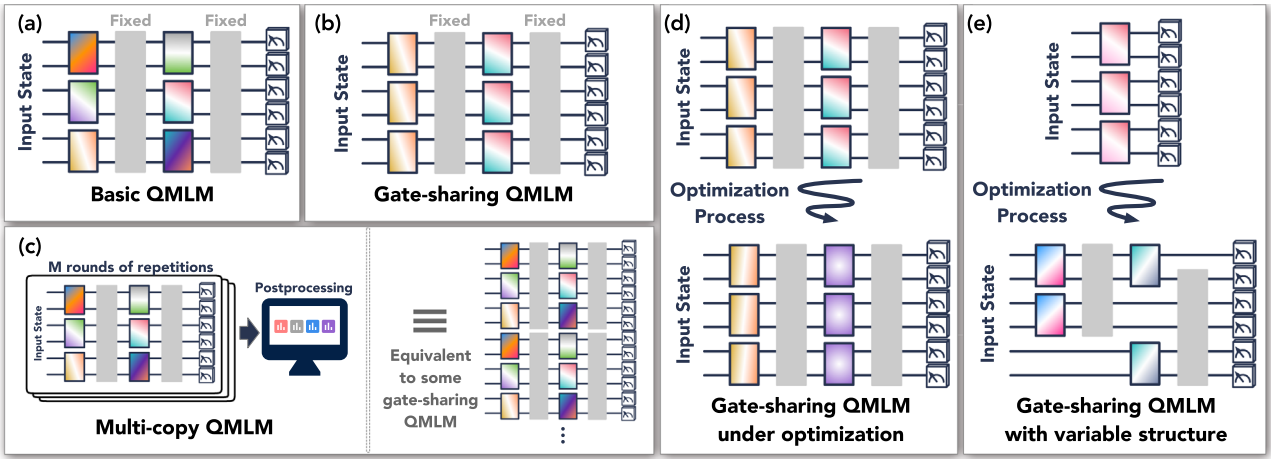


Fig. 1 | Various types of Quantum Machine Learning Models (QMLMs). Panel (**a**) depicts a basic QMLM with $T = 6$ independently parameterized gates. The gray boxes illustrate some global evolutions that are not trainable. Panel (**b**) shows a gate-sharing QMLM with $T = 2$ independently parameterized gates, each gate is repeatedly used for $M = 3$ times. In panel (**c**), we depict a multi-copy QMLM. We take measurement data from M rounds of a basic QMLM with $T = 6$ parameterized gates and post-process the measurement outcomes to produce an output. Running M copies of a basic QMLM with T gates is equivalent to running a gate-sharing QMLM with $T = 6$ parameterized gates, in which each gate is repeated M times. Panel (**d**) describes a gate-sharing QMLM under optimization. The parameterized gate to the left undergoes a small change, while the one to the right undergoes a large change. If we sort the changes Δ_1, Δ_2 from large to small, then $\Delta_1 \gg \Delta_2 \approx 0$. Finally, panel (**e**) illustrates gate-sharing QMLM with variable structure. The number T of parameterized gates changes throughout the optimization. The figure begins with $T = 1$ and ends with $T = 2$.

Figure 1 gives an overview of the different scenarios considered in this work. We begin with the basic form of our result. We consider a QMLM that has arbitrarily many non-trainable global quantum gates and T trainable local quantum gates. Here, by local we mean κ-local for some n-independent locality parameter κ, and a local quantum gate can be a unitary or a quantum channel acting on κ qubits. Then we have the following bound on the generalization error for the QMLM with final parameter setting $\boldsymbol{\alpha}^*$ after training:

Theorem 1. (*Basic QMLM*). For a QMLM with T parameterized local quantum channels, with high probability over training data of size N, we have that

$$\mathrm{gen}\,(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(\sqrt{\frac{T\log T}{N}}\right). \tag{5}$$

Remark 1. Theorem 1 directly implies sample complexity bounds: For any $\varepsilon > 0$, we can, with high success probability, guarantee that $\mathrm{gen}(\boldsymbol{\alpha}^*) \leqslant \varepsilon$, already with training data of size $N \sim T\log T/\varepsilon^2$, which scales effectively linearly with T, the number of parameterized gates.

For efficiently implementable QMLMs with $T \in \mathcal{O}(\mathrm{poly}n)$, a sample size of $N \in \mathcal{O}(\mathrm{poly}n/\varepsilon^2)$ is already sufficient. More concretely, if $T \in \mathcal{O}(n^D)$ for some degree D, then the corresponding sufficient sample complexity obtained from Theorem 1 satisfies $N \in \tilde{\mathcal{O}}(n^D/\varepsilon^2)$, where the $\tilde{\mathcal{O}}$ hides factors logarithmic in n. In the NISQ era[56], we expect the number T of trainable maps to only grow mildly with the number of qubits, e.g., as in the architectures discussed in refs. 18, 45, 57. In this case, Theorem 1 gives an especially strong guarantee.

In various QMLMs, such as QCNNs, the same parameterized local gates are applied repeatedly. One could also consider running the same QMLM multiple times to gather measurement data and then post-processing that data. In both cases, one should consider the QMLM as using the same parameterized local gates repeatedly. We assume each gate to be repeated at most M times. A direct application of Theorem 1 would suggest that we need a training data size N of roughly MT, the total number of parameterized gates. However, the required number of training data actually is much smaller:

Theorem 2. (*Gate-sharing QMLM*). Consider a QMLM with T independently parameterized local quantum channels, where each channel is reused at most M times. With high probability over training data of size N, we have

$$\mathrm{gen}\,(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(\sqrt{\frac{T\log(MT)}{N}}\right). \tag{6}$$

Thus, good generalization, as in Remark 1, can already be guaranteed, with high probability, when the data size effectively scales linearly in T (the number of independently parameterized gates) and only logarithmically in M (the number of uses). In particular, applying multiple copies of the QMLM in parallel does not significantly worsen the generalization performance compared to a single copy. Thus, as we discuss in Supplementary Note 3, Theorem 2 ensures that we can increase the number of shots used to estimate expectation values at the QMLM output without substantially harming the generalization behavior.

The optimization process of the QMLM also plays an important role in the generalization performance. Suppose that during the optimization process, the t^{th} local gate changed by a distance Δ_t. We can bound the generalization error by a function of the changes $\{\Delta_t\}_t$.

Theorem 3. (*Gate-sharing QMLM under optimization*). Consider a QMLM with T independently parameterized local quantum channels, where the t^{th} channel is reused at most M times and is changed by Δ_t during the optimization. Assume $\Delta_1 \geq \ldots \geq \Delta_T$. With high probability over training data of size N, we have

$$\mathrm{gen}\,(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(\min_{K=0,\ldots,T}\left\{\sqrt{\frac{K\log(MT)}{N}} + \sum_{k=K+1}^{T} M\Delta_k\right\}\right). \tag{7}$$

When only $K \ll T$ local quantum gates have undergone a significant change, then the generalization error will scale at worst linearly with K and logarithmically in the total number of parameterized gates MT. Given that recent numerical results suggest that the parameters in a deep parameterized quantum circuit only change by a small amount during training[58,59], Theorem 3 may find application in studying the generalization behavior of deep QMLMs.

Finally, we consider a more advanced type of variable ansatz optimization strategy that is also adopted in practice[60–63]. Instead of

https://doi.org/10.1038/s41467-022-32550-3

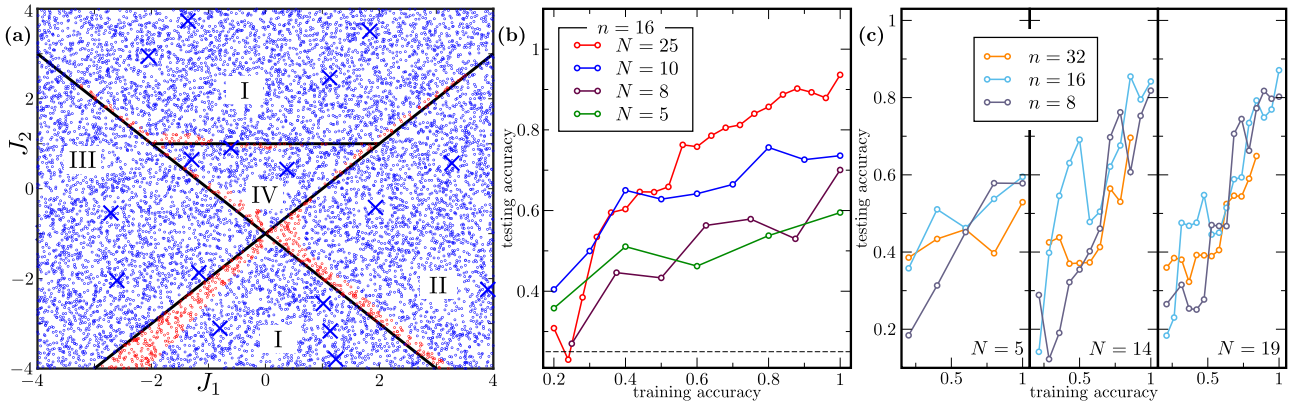


Fig. 2 | Generalization performance of quantum phase recognition. We employed the QCNN architecture for quantum phase recognition on ground states of the generalized cluster Hamiltonian H of Eq. (9). We evaluated the phase assigned by the QCNN to a point in the J_1-J_2-plane by sampling 8192 computational basis measurement outcomes and taking the least frequent outcome as the predicted phase. Panel (**a**) visualizes the performance of the QCNN for 16-qubits, trained with 30 data points, which were labelled according to the analytically determined phase diagram. Blue crosses denote training data points (not all 30 are shown). Blue (red) circles represent correctly (incorrectly) classified points. Panel (**b**) shows that, as the training data size increases, the training accuracy quickly becomes a good predictor for the testing accuracy on 10,000 randomly sampled points, i.e., the dependence of testing accuracy on training accuracy is approximately linear with slope increasing with N. The different points in the plot correspond to different parameter settings in the QCNN throughout the optimization. The dotted gray line shows the baseline accuracy of 25% achieved by random guessing. Panel (**c**) shows that the improvement in the slope with growing training data size is similar for different numbers of qubits, reflecting the at-worst polylogarithmic dependence of N on n predicted by our bounds.

fixing the structure of the QMLM, such as the number of parameterized gates and how the parameterized gates are interleaved with the fixed gates, the optimization algorithm could vary the structure, e.g., by adding or deleting parameterized gates. We assume that for each number T of parameterized gates, there are G_T different QMLM architectures.

Theorem 4. (*Gate-sharing QMLM with variable structure*). Consider a QMLM with an arbitrary number of parameterized local quantum channels, where for each $T > 0$, we have G_T different QMLM architectures with T parameterized gates. Suppose that after optimizing on the data, the QMLM has T independently parameterized local quantum channels, each repeated at most M times. Then, with high probability over input training data of size N,

$$\mathrm{gen}\,(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(\sqrt{\frac{T\log(MT)}{N}} + \sqrt{\frac{\log(G_T)}{N}}\right). \tag{8}$$

Thus, even if the QMLM can in principle use exponentially many parameterized gates, we can control the generalization error in terms of the number of parameterized gates used in the QMLM after optimization, and the dependence on the number of different architectures is only logarithmic. This logarithmic dependence is crucial as even in the cases when G_T grows exponentially with T, we have $\log(G_T)/N \in \mathcal{O}(T/N)$.

Numerical results

In this section we present generalization error results obtained by simulating the following two QML implementations: (1) using a QCNN to classify states belonging to different quantum phases, and (2) training a parameterized quantum circuit to compile a quantum Fourier transform matrix.

We begin with the quantum phase classification application. The QCNN architecture introduced in[45] generalizes the model of (classical) convolutional neural networks with the goal of performing pattern recognition on quantum data. It is composed of so-called *convolutional* and *pooling* layers, which alternate. In a convolutional layer, a sequence of translationally invariant parameterized unitaries on neighbouring qubits is applied in parallel, which works as a filter between feature maps in different layers of the QCNN. Then, in the pooling layers, a subset of the qubits are measured to reduce the dimensionality of the state while preserving the relevant features of the data. Conditioned on the corresponding measurement outcomes, translationally invariant parameterized 1-qubit unitaries are applied. The QCNN architecture has been employed for supervised QML tasks of classification of phases of matter and to devise quantum error correction schemes[45]. Moreover, QCNNs have been shown not to exhibit barren plateaus, making them a generically trainable QML architecture[27].

The action of a QCNN can be considered as mapping an input state ρ_{in} to an output state ρ_{out} given as $\rho_{\mathrm{out}} = \mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QCNN}}(\rho_{\mathrm{in}})$. Then, given ρ_{out}, one measures the expectation value of a task-specific Hermitian operator.

In our implementation, we employ a QCNN to classify states belonging to different symmetry protected topological phases. Specifically, we consider the generalized cluster Hamiltonian

$$H = \sum_{j=1}^{n}\left(Z_j - J_1 X_j X_{j+1} - J_2 X_{j-1} Z_j X_{j+1}\right), \tag{9}$$

where Z_i (X_i) denote the Pauli z (x) operator acting on qubit i, and where J_1 and J_2 are tunable coupling coefficients. As proved in[64], and as schematically shown in Fig. 2, the ground-state phase diagram of the Hamiltonian of Eq. (9) has four different phases: symmetry-protected topological (I), ferromagnetic (II), anti-ferromagnetic (III), and trivial (IV). In the Methods section, we provide additional details regarding the classical simulation of the ground states of H.

By sampling parameters in the (J_1, J_2) plane, we create a training set $\{(|\psi_i\rangle, y_i)\}_{i=1}^{N}$ composed of ground states $|\psi_i\rangle$ of H and their associated labels y_i. Here, the labels are in the form of length-two bit strings, i.e., $y_i \in \{0,1\}^2$, where each possible bit string corresponds to a phase that $|\psi_i\rangle$ can belong to. The QCNN maps the n-qubit input state $|\psi_i\rangle$ to a 2-qubit output state. We think of the information about the phase as being encoded into the output state by which of the 4 computational basis effect operators is assigned the smallest probability. Namely, we define the loss function as $\ell(\boldsymbol{\alpha}; |\psi_i\rangle, y_i) := \langle y_i | \mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QCNN}}(|\psi_i\rangle\langle\psi_i|) | y_i\rangle$. This

leads to an empirical risk given by

$$\hat{R}_S(\boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \langle y_i|\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QCNN}}(|\psi_i\rangle\langle\psi_i|)|y_i\rangle. \tag{10}$$

In Fig. 2, we visualize the phase classification performance achieved by our QCNN, trained according to this loss function, while additionally taking the number of misclassified points into account. Moreover, we show how the true risk, or rather the test accuracy as proxy for it, correlates well with the achieved training accuracy, already for small training data sizes. This is in agreement with our theoretical predictions, discussed in more detail in Supplementary Note 4, which for QCNNs gives a generalization error bound poly-logarithmic in the number of qubits. We note that refs. 65, 66 observed similarly favorable training data requirements for a related task of learning phase diagrams.

Next, we turn our attention to the unitary compiling application. Compiling is the task of transforming a high-level algorithm into a low-level code that can be implemented on a device. Unitary compiling is a paradigmatic task in the NISQ era where a target unitary is compiled into a gate sequence that complies with NISQ device limitations, e.g., hardware-imposed connectivity and shallow depth to mitigate errors. Unitary compiling is crucial to the quantum computing industry, as it is essentially always performed prior to running an algorithm on a NISQ device, and various companies have their own commercial compilers[67,68]. Hence, any ability to accelerate unitary compiling could have industrial impact.

Here we consider the task of compiling the unitary U of the n-qubit Quantum Fourier Transform (QFT)[69] into a short-depth parameterized quantum circuit $V(\boldsymbol{\alpha})$. For $V(\boldsymbol{\alpha})$ we employ the VAns (Variable Ansatz) algorithm[62,70], which uses a machine learning protocol to iteratively grow a parameterized quantum circuit by placing and removing gates in a way that empirically leads to lower loss function values. Unlike traditional approaches that train just continuous parameters in a fixed structure circuit, VAns also trains discrete parameters, e.g., gate placement or type of gate, to explore the architecture hyperspace. In Supplementary Note 5, we apply our theoretical results in this compiling scenario.

The training set for compilation is of the form $\{|\psi_i\rangle, U|\psi_i\rangle\}_{i=1}^{N}$, consisting of input states $|\psi_i\rangle$ and output states obtained through the action of U. The $|\psi_i\rangle$ are drawn independently from an underlying data-generating distribution. In our numerics, we consider three such distributions: (1) random computational basis states, (2) random (non-orthogonal) low-entangled states, and (3) Haar random n-qubit states. Note that states in the first two distributions are easy to prepare on a quantum computer, whereas states from the last distribution become costly to prepare as n grows. As the goal is to train $V(\boldsymbol{\alpha})$ to match the action of U on the training set, we define the loss function as the squared trace distance between $U|\psi_i\rangle$ and $V(\boldsymbol{\alpha})|\psi_i\rangle$, i.e., $\ell(\alpha; |\psi_i\rangle, U|\psi_i\rangle) := ||U|\psi_i\rangle\langle\psi_i|U^\dagger - V(\boldsymbol{\alpha})|\psi_i\rangle\langle\psi_i|V(\boldsymbol{\alpha})^\dagger||_1^2$. This leads to the empirical risk

$$\hat{R}_S(\boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} ||U|\psi_i\rangle\langle\psi_i|U^\dagger - V(\boldsymbol{\alpha})|\psi_i\rangle\langle\psi_i|V(\boldsymbol{\alpha})^\dagger||_1^2, \tag{11}$$

where $||\cdot||_1$ indicates the trace norm.

Figure 3 shows our numerical results. As predicted by our analytical results, we can, with high success probability, accurately compile the QFT when training on a data set of size polynomial in the number of qubits. Our numerical investigation shows a linear scaling of the training requirements when training on random computational basis states. This better than the quadratic scaling implied by a direct application of our theory, which holds for any arbitrary data-generating distribution. Approximate implementations of QFT with a reduced number of gates[71], combined with our results, could help to further study this apparent gap theoretically. When training on Haar random states, our numerics suggest that an even smaller number of training data points is sufficient for good generalization: Up to $n = 9$ qubits, we generalize well from a constant number of training data points, independent of the system size.

Even more striking are our results when initializing close to the solution. In this case, as shown in Fig. 4, we find that two training data points suffice to obtain accurate generalization, which holds even up to a problem size of 40 qubits. Our theoretical results in Theorem 3 do predict reduced training requirements when initializing near the solution. Hence, the numerics are in agreement with the theory, although they paint an even more optimistic picture and suggest that further investigation is needed to understand why the training data requirements are so low. While the assumption of initialization near the solution is only viable assuming additional prior knowledge, it could be justified in certain scenarios. For example, if the unitaries to be compiled depend on a parameter, e.g., time, and if we have already compiled the unitary for one parameter setting, we might use this as initialization for unitaries with a similar parameter.

Discussion

We conclude by discussing the impact of our work on specific applications, a comparison to prior work, the interpretation of our results from the perspective of quantum advantage, and some open questions.

We begin with a discussion of the impact on specific applications. Quantum phase classification is an exciting application of QML, to which Ref. 45 has successfully applied QCNNs. However, Ref. 45 only provided a heuristic explanation for the good generalization performance of QCNNs. Here, we have presented a rigorous theory that encompasses QCNNs and explains their performance, and we have confirmed it numerically for a fairly complicated phase diagram and a wide range of system sizes. In particular, our analysis allows us to go beyond the specific model of QCNNs and extract general principles for how to ensure good generalization. As generating training data for this problem asks an experimenter to prepare a variety of states from different phases of matter, which will require careful tuning of different parameters in the underlying Hamiltonian, good generalization guarantees for small training data sizes are crucial to allow for the implementation of phase classification through QML in actual physical experiments.

Several successful protocols for unitary compiling make use of training data[52–54]. However, prior work has relied on training data sets whose size scaled exponentially with the number of qubits. This scaling is problematic, both because it suggests a similarly bad scaling of the computational complexity of processing the data and because generating training data can be expensive in actual physical implementations. Our generalization bounds provide theoretical guarantees on the performance that unitary compiling with only polynomial-size training data can achieve, for the relevant case of efficiently implementable target unitaries. As we have numerically demonstrated in the case of the Quantum Fourier Transform, this significant reduction in training data size makes unitary compiling scalable beyond what previous approaches could achieve. Moreover, our results provide new insight into why the VAns algorithm[62] is successful for unitary compiling. We believe that the QML perspective on unitary compiling advocated for in this work will lead to new and improved ansätze, which could scale to even larger systems.

Recent methods for variational dynamical simulation rely on quantum compiling to compile a Trotterized unitary into a structured ansatz with the form of a diagonalization[48,49,72,73]. This technique allows for quantum simulations of times longer than an iterated Trotterization because parameters in the diagonalization may be changed by hand to provide longer-time simulations with a fixed depth circuit. We expect the quantum compiling results presented here to carry over to

https://doi.org/10.1038/s41467-022-32550-3

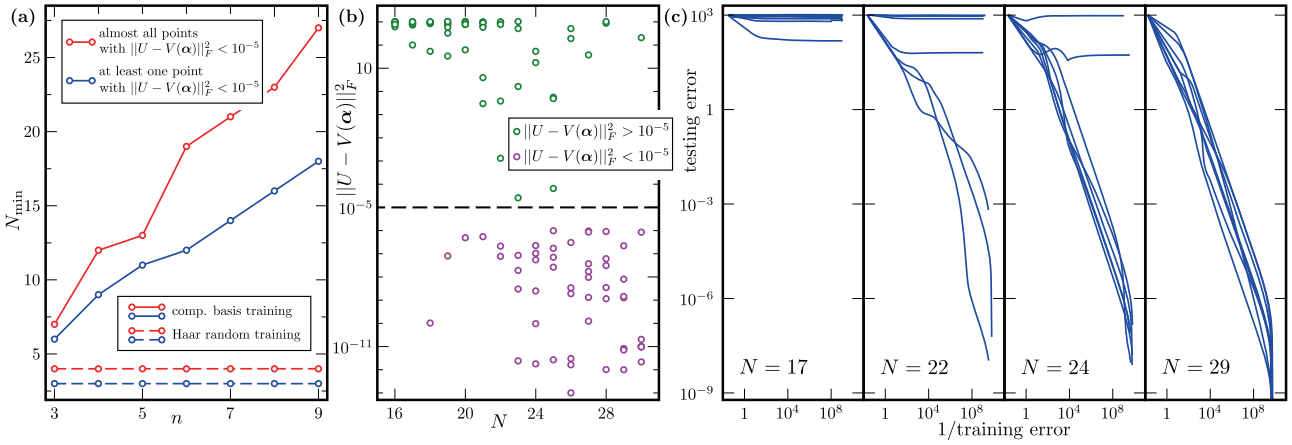


Fig. 3 | Generalization performance of variational unitary compiling. We employed a variable structure QMLM (as discussed near Theorem 4). Panel (**a**) shows the dependence of N_{min}, the minimum training data size for accurate compilation, on n, the number of qubits. Accurate compilation is defined as achieving $\| U - V(\boldsymbol{\alpha})\|_F^2 < 10^{-5}$ on 1 out of 8 (blue) or on 7 out of 8 (red) runs. For training data with random computational basis inputs (solid lines), N_{min} scales linearly in n. When training on examples with Haar random inputs (dashed lines), N_{min} is constant up to system size $n = 9$. In Panel (**b**), for $n = 9$ qubits, we plot the prediction error of successfully trained (training cost $< 10^{-8}$) runs for 8 training data sets with $N = 16$ to $N = 30$ random computational basis inputs. Panel (**c**) shows the dependence of the testing error on the reciprocal of the training error for different training data sizes, in the case of 9 qubits. Here, the data consisted of random computational basis states and the corresponding outputs. As N increases, small training error becomes a more reliable predictor for small testing error. Each subplot shows 8 different training runs, trained on different training data sets.

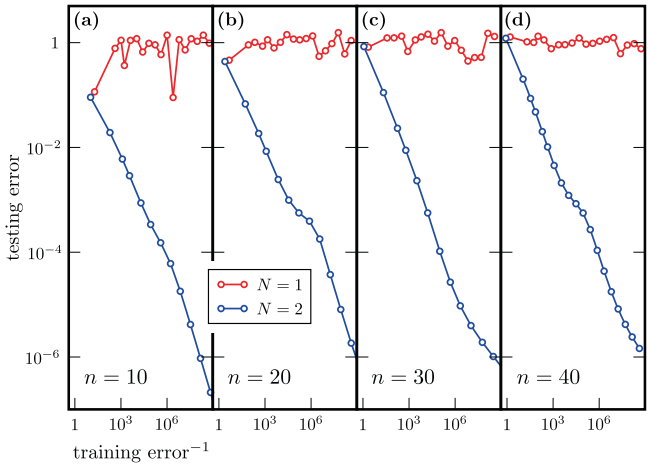


Fig. 4 | Performance of variational unitary compiling when initializing near the solution. Each panel shows the results of a single randomly initialized training run, where we used randomly drawn low-entangled states for training. The testing error on 20 test states, which we allow to be more strongly entangled than the states used during training, is plotted versus the reciprocal of the training error for training data sizes $N = 1, 2$, for different system sizes n. A training data set of size $N = 1$ is not sufficient to guarantee good generalization: Even with decreasing training error, the testing error remains large. In contrast, assuming favorably initialized training, $N = 2$ training data points suffice for good generalization, even for up to $n = 40$ qubits.

this application. This will allow these variational quantum simulation methods to use fewer training resources (either input-output pairs, or entangling auxiliary systems), yet still achieve good generalization and scalability.

Discovering quantum error correcting codes can be viewed as an optimization problem[44,45,74–78]. Furthermore, it can be thought of as a machine learning problem, since computing the average fidelity of the code involves training data (e.g., chosen from a 2-design[44]). Significant effort has been made to solve this problem on classical computers[74–78]. Such approaches can benefit from our generalization bounds, potentially leading to faster classical discovery of quantum codes. More recently, it was proposed to use near-term quantum computers to find such codes[44,45]. Again our bounds imply good generalization performance with small training data for this application, especially for QCNNs[45], due to their logarithmic number of parameters.

Finally, autoencoders and generative adversarial networks (GANs) have recently been generalized to the quantum setting[13,43,79,80]. Both employ training data, and hence our generalization bounds provide quantitative guidance for how much training data to employ in these applications. Moreover, our results can provide guidance for ansatz design. While there is no standard ansatz yet for quantum autoencoders or quantum GANs, ansätze with a minimal number of parameters will likely lead to the best generalization performance.

Next, we give a comparison to previously known results. Some prior works have studied the generalization capabilities of quantum models, among them the classical learning-theoretic approaches of[81–89]; the more geometric perspective of[17,18]; and the information-theoretic technique of[20,37]. Independently of this work, Ref. 38 also investigated covering numbers in QMLMs. However our bounds are stronger, significantly more general, and broader in scope. We give a detailed comparison of our results to related work in Supplementary Note 1.

To view our results in the context of the quest for quantum advantage, it is important to note that we do not prove a quantum advantage of quantum over classical machine learning. However, generalization bounds for QMLMs are necessary to understand their potential for quantum advantage. Namely, QMLMs can outperform classical methods, assuming both achieve small training error, only in scenarios in which QMLMs generalize well, but classical ML methods do not. We therefore consider our results a guide in the search for quantum advantage of QML: We need to identify a task in which QMLMs with few trainable gates achieve small training error, but classical models need substantially higher model complexity to achieve the same goal. Then, our bounds guarantee that the QMLM performs well also on unseen data, but we expect the classical model to generalize poorly due to the high model complexity.

We conclude with some open questions. For QMLMs with exponentially many independently trainable gates, our generalization error bounds scale exponentially with n, and hence we do not make nontrivial claims about this regime. However, this does not yet imply that exponential-size QMLMs have bad generalization behavior. Whether and under which circumstances this is indeed the case is an interesting

open question (e.g., see[17,37] for some initial results). More generally, one can ask: Under what circumstances will a QMLM, even one of polynomial size, outperform our general bound. For example, if we have further prior knowledge about the loss, arising from specific target applications, it might be possible to use this information to tighten our generalization bounds. Moreover, as our generalization bounds are valid for arbitrary data-generating distributions, they may be overly pessimistic for favorable distributions. Concretely, in our numerical experiments for unitary compiling, highly entangled states were more favorable than especially efficiently preparable states from the perspective of generalization. It may thus be interesting to investigate distribution-specific tightenings of our results. Finally, it may be fruitful to combine the generalization bounds for QMLMs studied in this work and the effect of data encodings in[86] to yield a better picture on generalization in quantum machine learning.

Methods

This section gives an overview over our techniques. First, we outline the proof strategy that leads to the different generalization bounds stated above. Second, we present more details about our numerical investigations.

Analytical methods

An established approach to generalization bounds in classical statistical learning theory is to bound a complexity measure for the class under consideration. Metric entropies, i.e., logarithms of covering numbers, quantify complexity in exactly the way needed for generalization bounds, as one can show using the chaining technique from the theory of random processes[90,91]. Therefore, a high level view of our proof strategy is: We establish novel metric entropy bounds for QMLMs and then combine these with known generalization results from classical learning theory. The strongest form of our generalization bounds is the following.

Theorem 5. (*Mother theorem*). Consider a QMLM with an arbitrary number of parameterized local quantum channels, where for each $T > 0$, we have G_T different QMLM architectures with T trainable local gates. Suppose that after optimizing on the training data, the QMLM has T independently parameterized local quantum channels, where the t^{th} channel is reused at most M times and is changed by Δ_t during the optimization. Without loss of generality, assume $\Delta_1 \geq \ldots \geq \Delta_T$. Then with high probability over input training data of size N, we have

$$\operatorname{gen}(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(\min_{K=0,\ldots,T} f(K) + \sqrt{\frac{\log(G_T)}{N}} \right), \tag{12}$$

where $f(K) := \sqrt{\frac{K\log(MT)}{N}} + \sum_{k=K+1}^{T} M\Delta_k$.

We give a detailed proof in Supplementary Note 3. There, we also describe a variant in case the loss function cannot be evaluated exactly, but only estimated statistically. Here, we present only a sketch of how to prove Theorem 5.

Before the proof sketch, however, we discuss how Theorem 5 relates to the generalization bounds stated above. In particular, we demonstrate how to obtain Theorems 1, 2, 3, and 4 as special cases of Theorem 5.

In the scenario of Theorem 1, the QMLM architecture is fixed in advance, each trainable map is only used once, and we do not take properties of the optimization procedure into account. In the language of Theorem 5, this means: There exists a single $T > 0$ with $G_T = 1$ and we have $G_{\tilde{T}} = 0$ for all $\tilde{T} \neq T$. Also, $M = 1$. And instead of taking the minimum over $K = 1, \ldots, T$, we consider the bound for $K = T$. Plugging this into the generalization bound of Theorem 5, we recover Theorem 1.

Similarly, Theorem 5 implies Theorems 2, 3, and 4. Namely, if we take $G_T = 1$ and $G_{\tilde{T}} = 0$ for all $\tilde{T} \neq T$, and evaluate the bound for $K = T$, we recover Theorem 2. Choosing $G_T = 1$ and $G_{\tilde{T}} = 0$ for all $\tilde{T} \neq T$, the bound of Theorem 5 becomes that of Theorem 3. Finally, we can obtain Theorem 4 by bounding the minimum in Theorem 5 in terms of the expression evaluated at $K = T$.

Now that we have established that Theorem 5 indeed implies generalization bounds for all the different scenarios depicted in Fig. 1, we outline its proof. The first central ingredient to our reasoning are metric entropy bounds for the class of all n-qubit CPTP maps that a QMLM as described in Theorem 5 can implement, where the distance between such maps is measured in terms of the diamond norm. Note: The trivial metric entropy bound obtained by considering this class of maps as compact subset of an Euclidean space of dimension exponential in n is not sufficient for our purposes since it scales exponentially in n. Instead, we exploit the layer structure of QMLMs to obtain a better bound. More precisely, we show: If we fix a QMLM architecture with T trainable 2-qubit maps and a number of maps $0 \leqslant K \leqslant T$, and we assume (data-dependent) optimization distances $\Delta_1 \geqslant \ldots \geqslant \Delta_T$, then it suffices to take (ε/KM)-covering nets for each of the sets of admissible 2-qubit CPTP maps for the first K trainable maps to obtain a $(\varepsilon + \sum_{k=K+1}^{T} M\Delta_k)$-covering net for the whole QMLM. The cardinality of a covering net built in this way, crucially, is independent of n, but depends instead on K, M, and T. In detail, its logarithm can effectively be bounded as $\in \mathcal{O}(K\log(MT/\varepsilon))$. This argument directly extends from the 2-local to the κ-local case, as we describe in Supplementary Note 3.

Now we employ the second core ingredient of our proof strategy. Namely, we combine a known upper bound on the generalization error in terms of the expected supremum of a certain random process with the so-called chaining technique. This leads to a generalization error bound in terms of a metric entropy integral. As we need a non-standard version of this bound, we provide a complete derivation for this strengthened form. This then tells us that, for each fixed T, M, K, and $\Delta_1 \geqslant \ldots \geqslant \Delta_T$, using the covering net constructed above, we can bound the generalization error as $\operatorname{gen}(\boldsymbol{\alpha}^*) \in \mathcal{O}(\sqrt{K\log(MT)/N} + \sum_{k=K+1}^{T} M\Delta_k)$, with high probability.

The last step of the proof consists of two applications of the union bound. The first instance is a union bound over the possible values of K. This leads to a generalization error bound in which we minimize over $K = 0, \ldots, T$. So far, however, the bound still applies only to any QMLM with fixed architecture. We extend it to variable QMLM architectures by taking a second union bound over all admissible numbers of trainable gates T and the corresponding G_T architectures. As this is, in general, a union bound over countably many events, we have to ensure that the corresponding failure probabilities are summable. Thus, we invoke our fixed-architecture generalization error bound for a success probability that is proportional to $(G_T T^2)^{-1}$. In that way, the union bound over all possible architectures yields the logarithmic dependence on G_T in the final bound and completes the proof of Theorem 5.

Numerical methods

This section discusses numerical methods used throughout the paper. The subsections give details on computational techniques applied to phase classification of the cluster Hamiltonian in Eq. (9) and Quantum Fourier Transform compilation.

Phase classification. The training and testing sets consist of ground states $|\psi_i\rangle$ of the cluster Hamiltonian in Eq. (9), computed for different coupling strengths (J_1, J_2). The states $|\psi_i\rangle$ were obtained with the translation invariant Density Matrix Renormalization Group[92]. The states in the training set (represented by blue crosses in Fig. 2a) are chosen to be away from phase transition lines, so accurate description of the ground states is already achieved at small bond dimension χ. That value determines the cost of further computation involving the states $|\psi_i\rangle$ and we keep it small for efficient simulation.

We use Matrix Product State techniques[93] to compute and optimize the empirical risk in Eq. (10). The main part of that calculation is the simulation of the action of the QCNN $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QCNN}}$ on a given ground state $|\psi_i\rangle$. The map $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QCNN}}$ consists of alternating convolutional and pooling layers. In our implementation the layers are translationally invariant and are represented by parameterized two-qubit gates. The action of a convolutional layer on an MPS amounts to updating two nearest neighbor MPS tensors in a way similar to the time-evolving block decimation algorithm[94]. The pooling layer is simulated in two steps. First, we simulate the action of all two-qubit gates on an MPS. This is analogous to the action of a convolutional layer, but performed on a different pair of nearest neighbor MPS tensors. This step is followed by a measurement of half of the qubits. We use the fact that the MPS can be written as a unitary tensor network and hence allows for perfect sampling techniques[95]. The measurement step results in a reduction of the system size by a factor of two.

We repeat the application of convolutional and pooling layers using the MPS as described above until the system size becomes small enough to allow for an exact description. A few final layers are simulated in a standard way and the empirical risk is given by a two-qubit measurement according to the label y_i, as in Eq. (10). The empirical risk is optimized with the Simultaneous Perturbation Stochastic Approximation algorithm[96]. We grow the number of shots used in pooling layer measurements as the empirical risk is minimized. This results in a shot-frugal optimization[97], as one can control the accuracy of the gradient based on the current optimization landscape.

Unitary compiling. In the Numerical results section, we show that the task of unitary compilation can be translated into minimization of the empirical risk $\hat{R}_S(\boldsymbol{\alpha})$ defined in Eq. (11). Here, $\boldsymbol{\alpha}=(\boldsymbol{\theta},\boldsymbol{k})$ denotes a set of parameters that specifies a trainable unitary $V(\boldsymbol{\alpha})$. The optimization is performed in the space of all shallow circuits. It has discrete and continuous components. The discrete parameters $\boldsymbol{k}$ control the circuit layout, that is, the placement of all gates used in the circuit. Those gates are described by the continuous parameters $\boldsymbol{\theta}$. The optimization $\min_{\boldsymbol{\alpha}} \hat{R}_S(\boldsymbol{\alpha})$ is performed with the recently introduced VAns algorithm[62,70]. The unitary $V(\boldsymbol{\alpha})$ is initialized with a circuit that consists of a few randomly placed gates. In subsequent iterations, VAns modifies the structure parameter $\boldsymbol{k}$ according to certain rules that involve randomly placing a resolution of the identity and removing gates that do not significantly contribute to the minimization of the empirical risk $\hat{R}_S(\boldsymbol{\alpha})$. The qFactor algorithm[54], modified to work with a set of pairs of states as opposed to a target unitary, is used to optimize over continuous parameters $\boldsymbol{\theta}$ for fixed $\boldsymbol{k}$. This optimization is performed after each update to the structure parameter $\boldsymbol{k}$. In subsequent iterations, VAns makes a probabilistic decision whether the new set of parameters α' is kept or rejected. This decision is based on the change in empirical risk $\hat{R}_S(\alpha')-\hat{R}_S(\boldsymbol{\alpha})$, an artificial temperature T, and a factor Λ that sets the penalty for growing the circuit too quickly. To that end, we employ a simulated annealing technique, gradually decreasing T and Λ, and repeat the iterations described above until $\hat{R}_S(\boldsymbol{\alpha})$ reaches a sufficiently small value.

Let us now discuss the methods used to optimize the empirical risk when $V(\boldsymbol{\alpha})$ is initialized close to the solution. Here, we start with a textbook circuit for performing the QFT and modify it in the following way. First, the circuit is rewritten such that it consists of two-qubit gates only. Next, each two-qubit gate u is replaced with $u'=ue^{i\delta h}$, where h is a random Hermitian matrix and δ is chosen such that $||u-u'||=\epsilon$ for an initially specified ϵ. The results presented in the Numerical results section are obtained with $\epsilon=0.1$. The perturbation considered here does not affect the circuit layout and hence the optimization over continuous parameters $\boldsymbol{\theta}$ is sufficient to minimize the empirical risk $\hat{R}_S(\boldsymbol{\alpha})$. We use qFactor to perform that optimization.

The input states $|\psi_i\rangle$ in the training set $\{|\psi_i\rangle, U_{\mathrm{QFT}}|\psi_i\rangle\}_{i=1}^{N}$ are random MPSs of bond dimension $\chi=2$. The QFT is efficiently simulable[98] for such input states, which means that $U_{\mathrm{QFT}}|\psi_i\rangle$ admits an efficient MPS description. Indeed, we find that a bond dimension $\chi<20$ is sufficient to accurately describe $U_{\mathrm{QFT}}|\psi_i\rangle$. In summary, the use of MPS techniques allows us to construct the training set efficiently. Note that the states $V(\boldsymbol{\alpha})|\psi_i\rangle$ are in general more entangled than $U_{\mathrm{QFT}}|\psi_i\rangle$, especially at the beginning of the optimization. Because of that, we truncate the evolved MPS during the optimization. We find that a maximal allowed bond dimension of 100 is large enough to perform stable, successful minimization of the empirical risk with qFactor. The testing is performed with 20 randomly chosen initial states. We test with bond dimension $\chi=10$ MPSs, so the testing is done with more strongly entangled states than the training. Additionally, for system sizes up to 16 qubits, we verify that the trained unitary V is close (in the trace norm) to U_{QFT}, when training is performed with at least two states.

Data availability

The data generated and analyzed during the current study are available from the authors upon request.

Code availability

Further implementation details are available from the authors upon request.

References

1. Vapnik, V. N. & Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Th. Prob. App.* **16**, 264–280 (1971).
2. Pollard, D. *Convergence of stochastic processes* (Springer, 1984).
3. Giné, E. & Zinn, J. Some limit theorems for empirical processes. *Ann. Probability* 929–989. https://doi.org/10.1214/aop/1176993138 (1984).
4. Dudley, R. M. *Uniform Central Limit Theorems* (Cambridge University Press, 1999).
5. Bartlett, P. L. & Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Machine Learning Res.* **3**, 463–482 (2002).
6. Biamonte, J. Quantum machine learning. *Nature* **549**, 195–202 (2017).
7. Schuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemporary Phys.* **56**, 172–185 (2015).
8. Schuld, M., Sinayskiy, I. & Petruccione, F. The quest for a quantum neural network. *Quantum Inf. Process.* **13**, 2567–2586 (2014).
9. Dunjko, V. & Briegel, H. J. Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Rep. Prog. Phys.* **81**, 074001 (2018).
10. Cerezo, M. Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625–644 (2021).
11. Havlíček, V. Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209–212 (2019).
12. Farhi, E. & Neven, H. *Classification with quantum neural networks on near term processors*, arXiv preprint arXiv:1802.06002 (2018).
13. Romero, J., Olson, J. P. & Aspuru-Guzik, A. Quantum autoencoders for efficient compression of quantum data. *Quantum Sci. Technol.* **2**, 045001 (2017).
14. Wan, K. H., Dahlsten, O., Kristjánsson, H., Gardner, R. & Kim, M. S. Quantum generalisation of feedforward neural networks. *npj Quantum Inf.* **3**, 1–8 (2017).
15. Larocca, M., Ju, N., García-Martín, D., Coles, P. J. & Cerezo, M. Theory of overparametrization in quantum neural networks, arXiv preprint arXiv:2109.11676 (2021).

16. Schatzki, L., Arrasmith, A., Coles, P. J. & Cerezo, M. Entangled datasets for quantum machine learning, arXiv preprint arXiv:2109.03400 (2021).
17. Huang, H. Y. Power of data in quantum machine learning. *Nat. Commun.* **12**, 1–9 (2021).
18. Abbas, A. The power of quantum neural networks. *Nat. Comput. Sci.* **1**, 403–409 (2021).
19. Liu, Y., Arunachalam, S. & Temme, K. A rigorous and robust quantum speed-up in supervised machine learning, https://doi.org/10.1038/s41567-021-01287-z *Nat. Phys.*, 1–5 (2021).
20. Huang, H. Y., Kueng, R. & Preskill, J. Information-theoretic bounds on quantum advantage in machine learning. *Phys. Rev. Lett.* **126**, 190505 (2021).
21. Aharonov, D., Cotler, J. & Qi, X. L. Quantum algorithmic measurement. *Nat. Commun.* **13**, 1–9 (2022).
22. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 1–6 (2018).
23. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nat. Commun.* **12**, 1–12 (2021).
24. Cerezo, M. & Coles, P. J. Higher order derivatives of quantum neural networks with barren plateaus. *Quantum Sci. Technol.* **6**, 035006 (2021).
25. Arrasmith, A., Cerezo, M., Czarnik, P., Cincio, L. & Coles, P. J. Effect of barren plateaus on gradient-free optimization. *Quantum* **5**, 558 (2021).
26. Holmes, Z., Sharma, K., Cerezo, M. & Coles, P. J. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum* **3**, 010313 (2022).
27. Pesah, A. Absence of barren plateaus in quantum convolutional neural networks. *Phys. Rev. X* **11**, 041011 (2021).
28. Volkoff, T. & Coles, P. J. Large gradients via correlation in random parameterized quantum circuits. *Quantum Sci. Technol.* **6**, 025008 (2021).
29. Sharma, K., Cerezo, M., Cincio, L. & Coles, P. J. Trainability of dissipative perceptron-based quantum neural networks. *Phys. Rev. Lett.* **128**, 180505 (2022).
30. Holmes, Z. Barren plateaus preclude learning scramblers. *Phys. Rev. Lett.* **126**, 190501 (2021).
31. Marrero, C. O., Kieferová, M. & Wiebe, N. Entanglement-induced barren plateaus. *PRX Quantum* **2**, 040316 (2021).
32. Uvarov, A. V. & Biamonte, J. D. On barren plateaus and cost function locality in variational quantum algorithms. *J. Phys. A: Math. Theor.* **54**, 245301 (2021).
33. Patti, T. L., Najafi, K., Gao, X. & Yelin, S. F. Entanglement devised barren plateau mitigation. *Phys. Rev. Res.* **3**, 033090 (2021).
34. Wang, S. Noise-induced barren plateaus in variational quantum algorithms. *Nat. Commun.* **12**, 1–11 (2021).
35. Larocca, M. et al. *Diagnosing barren plateaus with tools from quantum optimal control*, arXiv preprint arXiv:2105.14377 (2021).
36. Thanaslip, S., Wang, S., Nghiem, N. A., Coles, P. J. & Cerezo, M. *Subtleties in the trainability of quantum machine learning models*, arXiv preprint arXiv:2110.14753 (2021).
37. Banchi, L., Pereira, J. & Pirandola, S. Generalization in quantum machine learning: a quantum information standpoint. *PRX Quantum* **2**, 040321 (2021).
38. Du, Y., Tu, Z., Yuan, X. & Tao, D. Efficient measure for the expressivity of variational quantum algorithms. *Phys. Rev. Lett.* **128**, 080506 (2022).
39. Poland, K., Beer, K. & Osborne, T. J. *No free lunch for quantum machine learning*, arXiv preprint arXiv:2003.14103 (2020).
40. Sharma, K. Reformulation of the no-free-lunch theorem for entangled datasets. *Phys. Rev. Lett.* **128**, 070501 (2022).
41. Bharti, K. Noisy intermediate-scale quantum algorithms. *Rev. Modern Phys.* **94**, 015004 (2022).
42. Endo, S., Cai, Z., Benjamin, S. C. & Yuan, X. Hybrid quantum-classical algorithms and quantum error mitigation. *J. Phys. Soc. Japan* **90**, 032001 (2021).
43. Romero, J. & Aspuru-Guzik, A. Variational quantum generators: Generative adversarial quantum machine learning for continuous distributions. *Adv. Quantum Technol.* **4**, 2000003 (2021).
44. Johnson, P. D., Romero, J., Olson, J., Cao, Y. & Aspuru-Guzik, A. *Qvector: an algorithm for device-tailored quantum error correction*, arXiv preprint arXiv:1711.02249 (2017).
45. Cong, I., Choi, S. & Lukin, M. D. Quantum convolutional neural networks. *Nat. Phys.* **15**, 1273–1278 (2019).
46. Khatri, S. Quantum-assisted quantum compiling. *Quantum* **3**, 140 (2019).
47. Sharma, K., Khatri, S., Cerezo, M. & Coles, P. J. Noise resilience of variational quantum compiling. *New J. Phys.* **22**, 043006 (2020).
48. Cirstoiu, C. Variational fast forwarding for quantum simulation beyond the coherence time. *npj Quantum Inf.* **6**, 1–10 (2020).
49. Commeau, B. et al. *Variational hamiltonian diagonalization for dynamical quantum simulation*, arXiv preprint arXiv:2009.02559 (2020).
50. Endo, S., Sun, J., Li, Y., Benjamin, S. C. & Yuan, X. Variational quantum simulation of general processes. *Phys. Rev. Lett.* **125**, 010501 (2020).
51. Li, Y. & Benjamin, S. C. Efficient variational quantum simulator incorporating active error minimization. *Phys. Rev. X* **7**, 021050 (2017).
52. Cincio, L., Subaşí, Y., Sornborger, A. T. & Coles, P. J. Learning the quantum algorithm for state overlap. *New J. Phys.* **20**, 113022 (2018).
53. Cincio, L., Rudinger, K., Sarovar, M. & Coles, P. J. Machine learning of noise-resilient quantum circuits. *PRX Quantum* **2**, 010324 (2021).
54. Younis, E. & Cincio, L. https://github.com/BQSKit/qfactor Quantum Fast Circuit Optimizer (qFactor).
55. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. & Latorre, J. Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020).
56. Preskill, J. Quantum computing in the nisq era and beyond. *Quantum* **2**, 79 (2018).
57. Romero, J. Strategies for quantum computing molecular energies using the unitary coupled cluster ansatz. *Quantum Sci. Technol.* **4**, 014008 (2018).
58. Shirai, N., Kubo, K., Mitarai, K. & Fuji, K. *Quantum tangent kernel*, arXiv preprint arXiv:2111.02951 (2021).
59. Liu, J., Tacchino, F., Glick, J. R., Jiang, L. & Mezzacapo, A. *Representation learning via quantum neural tangent kernels*, arXiv preprint arXiv:2111.04225 (2021).
60. Grimsley, H. R., Economou, S. E., Barnes, E. & Mayhall, N. J. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nat. Commun.* **10**, 1–9 (2019).
61. Tang, H. L. qubit-adapt-vqe: an adaptive algorithm for constructing hardware-efficient ansätze on a quantum processor. *PRX Quantum* **2**, 020310 (2021).
62. Bilkis, M., Cerezo, M., Verdon, G., Coles, P. J. & Cincio, L. *A semi-agnostic ansatz with variable structure for quantum machine learning*, arXiv preprint arXiv:2103.06712 (2021).
63. Zhu, L. et al. An adaptive quantum approximate optimization algorithm for solving combinatorial problems on a quantum computer, arXiv preprint arXiv:2005.10258 (2020).
64. Verresen, R., Moessner, R. & Pollmann, F. One-dimensional symmetry protected topological phases and their transitions. *Phys. Rev. B* **96**, 165124 (2017).
65. Kottmann, K., Corboz, P., Lewenstein, M. & Acín, A. Unsupervised mapping of phase diagrams of 2d systems from infinite projected

entangled-pair states via deep anomaly detection. *Sci. Post Phys.* **11**, 025 (2021).
66. Kottmann, K., Metz, F., Fraxanet, J. & Baldelli, N. Variational quantum anomaly detection: Unsupervised mapping of phase diagrams on a physical quantum computer. *Phys. Rev. Res.* **3**, 043184 (2021).
67. Cross, A. W., Bishop, L. S., Smolin, J. A. & Gambetta, J. M. *Open quantum assembly language*, arXiv preprint arXiv:1707.03429 (2017).
68. Smith, R. S., Curtis, M. J. & Zeng, W. J. *A practical quantum instruction set architecture*, arXiv preprint arXiv:1608.03355 (2016).
69. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge University Press, 2000).
70. Bilkis, M. *An implementation of VAns: A semi-agnostic ansatz with variable structure for quantum machine learning*. https://github.com/matibilkis/qvans.
71. Nam, Y., Su, Y. & Maslov, D. Approximate quantum fourier transform with o (n log (n)) t gates. *NPJ Quantum Inf.* **6**, 1–6 (2020).
72. Gibbs, J. et al. *Long-time simulations with high fidelity on quantum hardware*, arXiv preprint arXiv:2102.04313 (2021).
73. Geller, M. R., Holmes, Z., Coles, P. J. & Sornborger, A. Experimental quantum learning of a spectral decomposition. *Phys. Rev. Res.* **3**, 033200 (2021).
74. Fletcher, A. S., Shor, P. W. & Win, M. Z. Channel-adapted quantum error correction for the amplitude damping channel. *IEEE Trans. Inf. Theory* **54**, 5705–5718 (2008).
75. Fletcher, A. S., Shor, P. W. & Win, M. Z. Structured near-optimal channel-adapted quantum error correction. *Phys. Rev. A* **77**, 012320 (2008).
76. Kosut, R. L., Shabani, A. & Lidar, D. A. Robust quantum error correction via convex optimization. *Phys. Rev. Lett.* **100**, 020502 (2008).
77. Kosut, R. L. & Lidar, D. A. Quantum error correction via convex optimization. *Quantum Inf. Process.* **8**, 443–459 (2009).
78. Taghavi, S., Kosut, R. L. & Lidar, D. A. Channel-optimized quantum error correction. *IEEE Trans. Inf. Theory* **56**, 1461–1473 (2010).
79. Lloyd, S. & Weedbrook, C. Quantum generative adversarial learning. *Phys. Rev. Lett.* **121**, 040502 (2018).
80. Dallaire-Demers, P. L. & Killoran, N. Quantum generative adversarial networks. *Phys. Rev. A* **98**, 012324 (2018).
81. Caro, M. C. & Datta, I. Pseudo-dimension of quantum circuits. *Quantum Mach. Intell.* **2**, 14 (2020).
82. Bu, K., Koh, D. E., Li, L., Luo, Q. & Zhang, Y. Statistical complexity of quantum circuits. *Phys. Rev. A* **105**, 062431 (2022).
83. Bu, K., Koh, D. E., Li, L., Luo, Q. & Zhang, Y. *Effects of quantum resources on the statistical complexity of quantum circuits*, arXiv preprint arXiv:2102.03282 (2021).
84. Bu, K., Koh, D. E., Li, L., Luo, Q. & Zhang, Y. *Rademacher complexity of noisy quantum circuits*, arXiv preprint arXiv:2103.03139 (2021).
85. Gyurik, C., van Vreumingen, D. & Dunjko, V. *Structural risk minimization for quantum linear classifiers*, arXiv preprint arXiv:2105.05566 (2021).
86. Caro, M. C., Gil-Fuster, E., Meyer, J. J., Eisert, J. & Sweke, R. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum* **5**, 582 (2021).
87. Chen, C. C. On the expressibility and overfitting of quantum circuit learning. *ACM Trans. Quantum Comput.* **2**, 1–24 (2021).
88. Popescu, C. M. Learning bounds for quantum circuits in the agnostic setting. *Quantum Inf. Process.* **20**, 1–24 (2021).
89. Cai, H., Ye, Q. & Deng, D. L. Sample complexity of learning parametric quantum circuits. *Quantum Sci. Technol.* **7**, 025014 (2022).
90. Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning* (MIT Press, 2018).
91. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge University Press, 2018).
92. White, S. R. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69**, 2863 (1992).
93. Orús, R. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. *Ann. Phys.* **349**, 117–158 (2014).
94. Vidal, G. Classical simulation of infinite-size quantum lattice systems in one spatial dimension. *Phys. Rev. Lett.* **98**, 070201 (2007).
95. Ferris, A. J. & Vidal, G. Perfect sampling with unitary tensor networks. *Phys. Rev. B* **85**, 165146 (2012).
96. Spall, J. C. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins Apl Technical Digest* **19**, 482–492 (1998).
97. Kübler, J. M., Arrasmith, A., Cincio, L. & Coles, P. J. An adaptive optimizer for measurement-frugal variational algorithms. *Quantum* **4**, 263 (2020).
98. Browne, D. E. Efficient classical simulation of the quantum fourier transform. *New J. Phys.* **9**, 146 (2007).

Acknowledgements

M.C.C. was supported by the TopMath Graduate Center of the TUM Graduate School at the Technical University of Munich, Germany, the TopMath Program at the Elite Network of Bavaria, and by a doctoral scholarship of the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes). H.H. is supported by the J. Yang & Family Foundation. M.C., K.S., and L.C. were initially supported by the LANL LDRD program under project number 20190065DR. M.C. was also supported by the Center for Nonlinear Studies at LANL. K.S. acknowledges support the Department of Defense. A.T.S. and P.J.C. acknowledge initial support from the LANL ASC Beyond Moore's Law project. P.J.C. and L.C. were also supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research, under the Accelerated Research in Quantum Computing (ARQC) program. L.C. also acknowledges support from LANL LDRD program under project number 20200022DR. A.T.S. was additionally supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Science Center. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

Author contributions

The project was conceived by M.C.C., H.-Y.H., and P.J.C. The manuscript was written by M.C.C., H.-Y.H., M.C., K.S., A.S., L.C., and P.J.C. Theoretical results were proved by M.C.C. and H.-Y.H. The applications were conceived by M.C.C., H.-Y.H., M.C., K.S., A.S., L.C., and P.J.C. Numerical implementations were performed by M.C. and L.C.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-32550-3.

Correspondence and requests for materials should be addressed to Matthias C. Caro.

Supplementary Information for "Generalization in quantum machine learning from few training data"

Supplementary Note 1. Related Work

1. Related Work on Generalization Bounds for Quantum Machine Learning

In statistical learning theory, a variety of techniques for obtaining generalization bounds are known. The classical approach is based on complexity measures for the class of functions describing the machine learning model (MLM) under consideration. Among these complexity measures, the VC-dimension [1], the pseudo-dimension [2], the Rademacher complexity [3, 4], and covering numbers (and the related metric entropies) [5] are particularly well known. More recently, different approaches that take properties of the learning algorithm into account have been investigated, such as stability (introduced by [6]), differential privacy (going back to [7]), sample compression (due to [8]), and the PAC-Bayesian framework (described in [9]). The theory of generalization for quantum machine learning (QML) is less developed. Nevertheless, there has been some prior work, of which we now give an overview.

Ref. [10] proved bounds on the pseudo-dimension of quantum circuits in which the local unitaries can be varied. In particular, these pseudo-dimension bounds imply generalization bounds for learning polynomial-depth unitary quantum circuits from data. While the data encoding considered in Ref. [10] was a simple product encoding, this can be understood as an early investigation of the generalization behavior of variational quantum circuits. In particular, the techniques of Ref. [10] can also be applied for more general quantum data encodings. Ref. [11] has recently extended the generalization guarantees of Ref. [10] from the realizable to the agnostic setting, using covering number arguments. We note that all our generalization bounds apply to the agnostic setting, but for more general QMLMs than considered in [11].

Ref. [12] suggested the so-called effective dimension, derived from the (empirical) Fisher information matrix, as a complexity measure for the parameter space of a QMLM. In particular, Ref. [12] showed how to derive generalization bounds from bounds on the effective dimension and investigated this complexity measure numerically for different QMLMs. Contrary to the conclusions drawn in Ref. [12], the recommendations for QMLMs which we deduce from our generalization bounds are not unequivocably in favour of higher expressivity. Instead, we emphasize the ability to fit training data and the ability to generalize have to be balanced carefully. See Section III. in the main text for a discussion of implications of our results for a potential quantum advantage in QML.

Refs. [13–15] studied the Rademacher complexity of parameterized quantum circuits and thus QMLMs. They proved bounds on this complexity measure that depend on the size and depth of the circuit as well as on a measure of magic in the circuit. Ref. [14] provides a resource-theoretic perspective on the Rademacher complexity of a quantum circuit and Ref. [15] investigated the effects of noise in the circuit.

We mention one more related work that approaches generalization in QML via complexity measures. Ref. [16] provides bounds on covering numbers of QMLMs and, using these, deduces generalization bounds. We have developed our approach independently from Ref. [16] and have obtained both stronger and more general results. In particular, Theorem 6 shows that the generalization error bound scales as $\sqrt{T/N}$, where T is the number of trainable gates and N is the number of training data, compared to $T/\sqrt{N}$ in Theorem 2 in the first version of Ref. [16]. In addition, and in contrast to Ref. [16], we also consider the practically relevant scenarios of CPTP (not just unitary) QMLMs, of multiple uses/copies of trainable maps, and of variable QMLM structure. Moreover, our optimization-dependent generalization bounds for QMLMs are the first bounds of this kind for QML and showcase a new way of using covering numbers.

Ref. [17] has proposed an information-theoretic strategy towards studying the approximation and generalization capabilities of QMLMs. In particular, Ref. [17] demonstrates how the approximation and generalization errors of a QMLM can be bounded in terms of (Rényi) mutual informations between the quantum embedding achieved by the QMLM (before the final measurement) and the label or instance marginals of the data, respectively.

Ref. [18] considered a class of QMLM (quantum kernels) that is equivalent to training arbitrarily deep quantum circuits. The work also established generalization error bounds to study when quantum machine learning models would predict more accurately than classical machine learning models. Ref. [18] showed that even if we are training an arbitrarily deep quantum circuits, the generalization performance can still be good if a certain geometric criterion is met. Ref. [19] provided generalization error bounds for quantum kernels in noisy quantum circuits, and Ref. [20] studied the generalization performance of quantum kernels for some embeddings. Our work considers finite size quantum circuits and the resulting generalization error bounds are very different.

Even more recently, Ref. [21] has proved bounds on the VC-dimension and the fat-shattering dimension of a QMLM, by viewing the QMLM in terms of a parameterized measurement performed on the quantum data encoding. These

complexity bounds lead to generalization bounds for QMLMs that depend on spectral properties (more precisely, rank or Frobenius norm) of the parameterized measurement.

Shortly thereafter, Ref. [22] studied the generalization capabilities of QMLMs with a focus on the strategy used to encode classical data into the quantum circuit. In particular, they considered data encodings via Hamiltonian evolutions, where data re-uploading is allowed. For corresponding QMLMs, Ref. [22] established generalization bounds that depend explicitly on properties of the Hamiltonians used for data-encoding. These results are complementary to our work: The generalization guarantees of Ref. [22] depend only on the encoding strategy used in the QMLM, whereas our results are in formulated in terms of properties of the trainable part of the QMLM only.

Ref. [23] investigated the expressibility and the generalization behavior of specific QMLMs. By combining light cone arguments with insights into how a specific data-encoding leads to effective dimensionality limitations (see also [22]), Ref. [23] obtained VC-dimension bounds for the hardware efficient ansatz. These bounds depend on the number of qubits and on the number of trainable layers. Ref. [23] interpreted the overall limitation on the VC-dimension imposed by the data-encoding as an automatic regularization, which is helpful in avoiding overfitting.

Lastly, Ref. [24] investigated a problem of learning parametrized unitary quantum circuits from training data consisting of pairs of input and corresponding output states. They established generalization bounds, and thus sample complexity bounds, by first identifying a universal family of variational quantum circuit architecture, then considering a finite discretization of this family, and finally applying a standard generalization bound for finite hypothesis classes. We note that the generalization guarantee obtained from Theorem 6 is tighter than that obtained in Ref. [24]: For a variational n-qubit QMLM with at most n^c gates, [24, Theorem 2] implies that a sample complexity of $\tilde{\mathcal{O}}(n^{c+1})$ suffices for good generalization, whereas Theorem 6 tells us that already $\tilde{\mathcal{O}}(n^c)$ samples suffice. Additionally, our generalization guarantees apply for more general architectures than those considered in [24].

2. Related Work on Quantum Phase Recognition

Recognizing quantum phases of matter is an important question in physics. Recently, many works have considered training machine learning models to classify quantum phases. The works include the use of quantum neural networks [25] and classical machine learning models [26–29]. Most of the existing works do not come with rigorous guarantees. Thus, it is not clear whether the respective machine learning models will predict well after training. Our work shows that when a quantum neural network, such as a QCNN [25], can perform well on a training set with a moderate amount of examples, the quantum neural network will also predict well on new data. This is particularly prominent in QCNNs, for which the required training data size scales at most polylogarithmically in the system size. However, in order for quantum neural networks to achieve a small training error, one still needs to address various challenges, such as barren plateau in the training landscape [30, 31].

Recently, [32] has proposed provably efficient classical machine learning models that can classify a wide range of quantum phases of matter, including symmetry-broken phases, topological phases, and symmetry-protected topological phases. These classical machine learning models are efficient in both computational time and the required training data [32]. Furthermore, the numerical experiments of [32] have shown that no labels of the different phases are needed to train the classical machine learning models. The classical algorithm can automatically uncover the quantum phases of matter in an unsupervised learning procedure.

It remains to be seen if QMLMs, such as QCNNs [25], can improve upon classical machine learning models in classifying quantum phases of matter. For example, [32] shows that the prediction performance of classical machine learning models sometimes degrades when the correlation length in the ground state wave function is high. It would be interesting to understand whether QMLMs can still work well when classical machine learning models fail.

3. Related Work on Quantum Compiling

Compiling of quantum circuits is a broad field with many distinct approaches. For example, temporal planning [33, 34], reinforcement learning [35], and supervised learning [36, 37] are three alternative approaches that have been applied to quantum compiling. Moreover, while classical methods for quantum compiling are the most common, it has also been proposed to do quantum-assisted quantum compiling where a quantum computer is involved in the compiling process [38–41].

While not all methods employ training data, it is worth noting that some state-of-the-art methods are in fact based on training data [36, 37, 42]. It is also worth remarking that noise-aware quantum compiling methods can involve training data [37]. For these methods, it has largely been assumed that one would need an amount of training data that grows exponentially with the number of qubits. Naturally, this exponential scaling places a cutoff on the size of

unitaries that one can compile. However, with our results in mind (allowing for only polynomial-sized training sets), this cutoff can be significantly extended to larger unitary sizes.

For quantum compiling, the benefit of our work is two-fold, in that both classical methods and quantum methods can potentially be sped-up. Classical methods for quantum compiling are currently being used in the quantum computing industry to enhance the performance of cloud-based quantum computing (e.g., by companies such as Rigetti and IBM). Therefore, speeding up classical methods for quantum compiling can potentially have a direct impact on cloud-based quantum computing. Both standard compiling and noise-aware compiling are important for industrial near-term quantum computing, and our work impacts both of these approaches.

In addition, quantum-assisted methods for quantum compiling can also reduce their resource costs based on our results. Variational quantum algorithms for quantum compiling have been introduced [38–41]. Specifically, Refs. [38, 39, 43] discussed methods that employ an entangled state on $2n$ qubits to compile an n-qubit unitary. Due to our work, this entangled state can apparently be reduced in size, namely only needing a Schmidt rank that is polynomial in n (instead of a Schmidt rank that is exponential in n). Ref. [40] proposed a slightly different approach that did not involve an auxiliary system, but simply used multiple training data points. Our work shows that the amount of training data here does not need to grow exponentially in n, making the approach in Ref. [40] potentially scalable.

Finally, we note that variable ansatz methods (e.g., Ref. [44, 45]) for quantum compiling is a state-of-the-art approach that is employed, e.g., in Refs. [36, 37]. As noted in the main text, our results are general enough to cover the variable ansatz case (where the structure of the circuit changes during the optimization). Hence we provide guidance for how much training data is needed for the variable ansatz case as well.

Supplementary Note 2. Auxiliary Lemmata

Before presenting our results, we use this section to recall some well known auxiliary results that enter our proofs.

1. Auxiliary Lemmata from Statistical Learning Theory

We use two standard concentration inequalities. The first is due to Wassily Hoeffding.

Lemma 1 (Hoeffding's Concentration Inequality [46])**.** Let $X_1, \ldots, X_N$ be independent $\mathbb{R}$-valued random variables. Assume that, for every $1 \leqslant i \leqslant N$, $X_i \in [a_i, b_i]$ almost surely, where $a_i, b_i \in \mathbb{R}$, $a_i \leqslant b_i$. Then, for every $\varepsilon > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geqslant \varepsilon\right] \leqslant \exp\left(-2\varepsilon^2 / \sum_{i=1}^{N}(b_i - a_i)^2\right), \tag{1}$$

$$\mathbb{P}\left[\left|\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i])\right| \geqslant \varepsilon\right] \leqslant 2\exp\left(-2\varepsilon^2 / \sum_{i=1}^{N}(b_i - a_i)^2\right). \tag{2}$$

The second is the bounded differences inequality, originally due to Colin McDiarmid.

Lemma 2 (McDiarmid's Concentration Inequality [47])**.** Let $X_1, \ldots, X_N$ be independent random variables, each with values in $\mathcal{Z}$. Let $\varphi : \mathcal{Z}^N \to \mathbb{R}$ be a measurable function s.t., whenever $z \in \mathcal{Z}^n$ and $z' \in \mathcal{Z}^n$ differ only in the i^{th} entry, then $|\varphi(z) - \varphi(z')| \leqslant b_i$. Then, for every $\varepsilon > 0$, we have

$$\mathbb{P}\left[\varphi(Z) - \mathbb{E}[\varphi(Z)] \geqslant \varepsilon\right] \leqslant \exp\left(-2\varepsilon^2 / \sum_{i=1}^{N} b_i^2\right). \tag{3}$$

The third well known ingredient that we will employ in our reasoning without giving a proof is the following.

Lemma 3 (Massart's Lemma [48])**.** Let $N \in \mathbb{N}$. Let $A \subset \mathbb{R}^N$ be a finite set contained in a Euclidean ball of radius $r > 0$. Then

$$\mathbb{E}_\sigma\left[\sup_{a \in A}\frac{1}{N}\sum_{i=1}^{N}\sigma_i a_i\right] \leqslant \frac{r\sqrt{2\log|A|}}{N}, \tag{4}$$

where the expectation is w.r.t. i.i.d. Rademacher random variables $\sigma_1, \ldots, \sigma_N$.

2. Auxiliary Lemmata from Quantum Information Theory

From quantum information theory, we crucially make use of the following lemma.

Lemma 4 (Subadditivity of diamond distance; see [49], Proposition 3.48)**.** For any completely positive and trace-preserving maps $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$, where $\mathcal{B}$ and $\mathcal{D}$ map from n-qubit to m-qubit systems and $\mathcal{A}$ and $\mathcal{C}$ map from m-qubit to k-qubit systems, we have the following inequality

$$\|\mathcal{A}\mathcal{B} - \mathcal{C}\mathcal{D}\|_\diamond \leqslant \|\mathcal{A} - \mathcal{C}\|_\diamond + \|\mathcal{B} - \mathcal{D}\|_\diamond. \tag{5}$$

Also, to translate between the spectral norm of unitaries and the diamond norm of the corresponding channels, we employ the following result.

Lemma 5 (Spectral norm and diamond norm of unitary channels)**.** Let $\mathcal{U}(\rho) = U\rho U^\dagger$ and $\mathcal{V}(\rho) = V\rho V^\dagger$ be unitary channels. Then, $\frac{1}{2}\|\mathcal{U}(|\psi\rangle\langle\psi|) - \mathcal{V}(|\psi\rangle\langle\psi|)\|_1 \leqslant \|(U-V)|\psi\rangle\|_{\ell_2}$ for any pure state $|\psi\rangle$. Therefore,

$$\tfrac{1}{2}\|\mathcal{U} - \mathcal{V}\|_\diamond \leqslant \|U - V\|. \tag{6}$$

Proof. The proof is adapted from [50]. Fix an input $|\psi\rangle$ and denote the output state vectors by $|u\rangle = U|\psi\rangle$ and $|v\rangle = V|\psi\rangle$, respectively. Normalization ensures that these state vectors obey $|\langle u, v\rangle| \leqslant 1$, as well as $\||u\rangle - |v\rangle\|_{\ell_2} = \sqrt{2(1 - \mathrm{Re}(\langle u, v\rangle))}$. Apply the Fuchs–van de Graaf relations [51] to convert the output trace distance into a (pure) output fidelity:

$$\begin{aligned}
\tfrac{1}{2}\||u\rangle\langle u| - |v\rangle\langle v|\|_1 &= \sqrt{1 - |\langle u, v\rangle|^2} && (7)\\
&= \sqrt{(1 + |\langle u, v\rangle|)(1 - |\langle u, v\rangle|)} && (8)\\
&\leqslant \sqrt{2(1 - \mathrm{Re}(\langle u, v\rangle))} && (9)\\
&= \||u\rangle - |v\rangle\|_{\ell_2}. && (10)
\end{aligned}$$

The diamond distance bound then is a direct consequence of this relation. Using the fact that stabilization is not necessary for computing the diamond distance of two unitary channels [49], we get

$$\begin{aligned}
\tfrac{1}{2}\|\mathcal{U} - \mathcal{V}\|_\diamond &= \max_{|\psi\rangle\langle\psi|} \tfrac{1}{2}\|\mathcal{U}(|\psi\rangle\langle\psi|) - \mathcal{V}(|\psi\rangle\langle\psi|)\|_1 && (11)\\
&\leqslant \max_{|\psi\rangle} \|(U - V)|\psi\rangle\|_{\ell_2} = \|U - V\|. && (12)
\end{aligned}$$

Here, we have also used the definition of the operator norm. □

Supplementary Note 3. Analytical Results: Details and Proofs

We first introduce some standard notation. Let $\mathcal{D}(\mathcal{H})$ denote the set of density operators (positive semi-definite with unit trace) acting on the Hilbert space $\mathcal{H}$. Let $\mathcal{L}(\mathcal{H})$ denote the space of square linear operators acting on $\mathcal{H}$. Let $\mathcal{L}(\mathcal{H}, \mathcal{H}')$ denote the set of linear operators taking $\mathcal{H}$ to a Hilbert space $\mathcal{H}'$. The trace norm of a linear operator $A \in \mathcal{L}(\mathcal{H}, \mathcal{H}')$ is defined as $\|A\|_1 := \mathrm{Tr}[|A|]$, where $|A| := \sqrt{A^\dagger A}$. The trace distance between any two operators $A, B \in \mathcal{L}(\mathcal{H}, \mathcal{H}')$ is $\|A - B\|_1$, and for two quantum states $\rho, \sigma \in \mathcal{D}(\mathcal{H})$ it is linearly related to the maximum success probability of distinguishing ρ and σ in a quantum hypothesis testing experiment. A linear map $\mathcal{N}_{A\to B} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is called a completely positive (CP) map if $(\mathcal{I}_R \otimes \mathcal{N}_{A\to B})(X_{RA})$ is positive semi-definite for all positive semi-definite $X_{RA} \in \mathcal{L}(\mathcal{H}_{RA})$, where $\mathcal{H}_{RA} = \mathcal{H}_R \otimes \mathcal{H}_A$ and the reference system R can be of arbitrary size. Moreover, a linear map $\mathcal{N}_{A\to B} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is trace preserving (TP) if $\mathrm{Tr}(\mathcal{N}_{A\to B}(X_A)) = \mathrm{Tr}(X_A)$ for all $X_A \in \mathcal{L}(\mathcal{H}_A)$. A linear map $\mathcal{N}_{A\to B}$ is called a quantum channel if it is completely positive and trace preserving (CPTP). Let $\mathcal{N}_{A\to B}$ and $\mathcal{M}_{A\to B}$ denote quantum channels. Then the diamond distance between $\mathcal{N}_{A\to B}$ and $\mathcal{M}_{A\to B}$ is defined as

$$\|\mathcal{N}_{A\to B} - \mathcal{M}_{A\to B}\|_\diamond := \sup_{\rho_{RA} \in \mathcal{D}(\mathcal{H}_{RA})} \|(\mathcal{I}_R \otimes \mathcal{N}_{A\to B})(\rho_{RA}) - (\mathcal{I}_R \otimes \mathcal{M}_{A\to B})(\rho_{RA})\|_1, \tag{13}$$

where $\mathcal{I}_R$ is the identity map acting on $\mathcal{H}_R$.

As a consequence of the convexity of the trace norm and the Schmidt decomposition theorem, it suffices to optimize Eq. (13) over pure states in $\mathcal{H}_{RA}$ with $\dim(\mathcal{H}_R) = \dim(\mathcal{H}_A)$.

With the notation in place, we now present our analytical results. Generalization performance depends crucially on the metric entropy, which characterizes both classical and quantum machine learning models [52]. Metric entropy is a measure of complexity or expressiveness for a set of objects endowed with a distance metric.

In Supplementary Note 3. 1., we take the diamond norm as the distance metric and prove metric entropy bounds for two sets of interest. First, we examine the set $\mathcal{U}_{\mathcal{A}}$ of all unitaries that can be represented using a (fixed) variational quantum circuit $\mathcal{A}$ with T parameterized 2-qubit unitary gates. More precisely, we consider the corresponding set of unitary channels. Second, we study the set $\mathcal{CPTP}_{\mathcal{A}}$ of all CPTP maps that can be represented using a (fixed) variational quantum circuit $\mathcal{A}$ with T parameterized 2-qubit CPTP maps. The latter scenario generalizes the former and corresponds to the difference between perfect and noisy implementations. Note that, in both cases, the variational quantum circuit itself could contain more than T gates. However, these additional gates would have to be fixed and not trainable.

Using these metric entropy bounds and variants thereof, we establish prediction error bounds for variational quantum machine learning models (QMLMs) in terms of the number of trainable elements in Supplementary Note 3. 2.. We consider different scenarios of interest, among them that of using multiple copies of a quantum neural network (such that parameters are reused over different copies), as well as both fixed and variable circuit architectures.

1. Covering Number Bounds for Variational Quantum Circuits

In this section, we provide bounds on the expressivity of the class of CPTP maps (or unitaries) that a quantum machine learning model (QMLM) can implement in terms of the number of trainable elements used in the architecture. As a measure of expressivity, we choose covering numbers and metric entropies w.r.t. (the metric induced by) the diamond norm. We first recall the general definition of covering numbers and metric entropies.

Definition 1 (Covering nets, covering numbers, and metric entropies)**.** Let (X, d) be a metric space. Let $K \subset X$ be a subset and let $\varepsilon > 0$.

- $N \subseteq K$ is an ε-covering net of K if $\forall x \in K\ \exists y \in N$ such that $d(x, y) \leqslant \varepsilon$. That is, $N \subseteq K$ is an ε-covering net of K if and only if K can be covered by ε-balls around the points in N.
- The covering number $\mathcal{N}(K, d, \varepsilon)$ is the smallest possible cardinality of an ε-covering net of K.
- The metric entropies $\log_2 \mathcal{N}(K, d, \varepsilon)$ are the logarithm of the covering numbers.

In finite-dimensional real spaces, the covering numbers of norm balls, and thereby of norm-bounded sets, can be bounded easily. We make use of this observation to provide basic covering number bounds for the classes of 2-qubit unitaries and 2-qubit CPTP maps. We first state the bound for the unitary case.

Lemma 6 (Covering number bounds for 2-qubit unitaries)**.** Let $\|\cdot\|$ be a unitarily invariant norm on complex 4×4-matrices. The covering number of the set of 2-qubit unitaries $\mathcal{U}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right)$ w.r.t. the norm $\|\cdot\|$ can be bounded as

$$\mathcal{N}\left(\mathcal{U}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right), \|\cdot\|, \varepsilon\right) \leqslant \left(\frac{6\|\mathbb{1}_{\mathbb{C}^4}\|}{\varepsilon}\right)^{32}, \quad \text{for } 0 < \varepsilon \leqslant \|\mathbb{1}_{\mathbb{C}^4}\|. \tag{14}$$

Proof. It is well known (see, e.g., Section 4.2 in [53]) that the covering numbers of a norm-ball of radius $R > 0$ around some point $x \in \mathbb{R}^K$, for $0 < \varepsilon \leqslant R$, can be bounded as

$$\mathcal{N}(B_R(x), \|\cdot\|, \varepsilon) \leqslant \left(1 + \frac{2R}{\varepsilon}\right)^K \leqslant \left(\frac{3R}{\varepsilon}\right)^K, \tag{15}$$

where the ball and the coverings are taken w.r.t. the same norm.

In our scenario, we can apply this as follows: As $\|\cdot\|$ is assumed to be unitarily invariant, we have $\|U\| = \|\mathbb{1}_{\mathbb{C}^4}\|$ for every unitary $U \in \mathcal{U}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right)$. In particular, we have, for $R := \|\mathbb{1}_{\mathbb{C}^4}\|$ that $\mathcal{U}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right) \subset B_R(0)$, where $B_R(0)$ is the ball of matrices with $4 \times 4 = 16$ complex entries around the 0-matrix is taken w.r.t. $\|\cdot\|$. Therefore, we have

$$\mathcal{N}\left(\mathcal{U}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right), \|\cdot\|, \varepsilon\right) \leqslant \mathcal{N}\left(B_R(0), \|\cdot\|, \frac{\varepsilon}{2}\right) \leqslant \left(\frac{6\|\mathbb{1}_{\mathbb{C}^4}\|}{\varepsilon}\right)^{2\cdot 16}, \quad \text{for } 0 < \varepsilon \leqslant \|\mathbb{1}_{\mathbb{C}^4}\|, \tag{16}$$

where the first step uses the approximate monotonicity of (interior) covering numbers (see, e.g., Section 4.2 in [53]). □

This covering number bound becomes particularly useful for the spectral norm, for which $\|\mathbb{1}_{\mathbb{C}^4}\| = 1$.
With an analogous reasoning, we can prove a covering number bound for 2-qubit CPTP maps.

Lemma 7 (Covering number bounds for 2-qubit CPTP maps)**.** The covering number of the set of 2-qubit CPTP maps $\mathcal{CPTP}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right)$ w.r.t. the diamond distance can be bounded as

$$\mathcal{N}\left(\mathcal{CPTP}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right), \|\cdot\|_\diamond, \varepsilon\right) \leqslant \left(\frac{6}{\varepsilon}\right)^{512}, \quad \text{for } 0 < \varepsilon \leqslant 1. \tag{17}$$

Proof. As CPTP maps have diamond norm equal to 1, this follows (analogously to the previous Lemma) by upper-bounding the covering number of the diamond-norm unit ball, which lives in a $(2^4 \times 2^4)$-dimensional space over the complex numbers. The latter can be achieved as in the previous Lemma. □

We combine these basic upper bounds for single trainable elements with sub-additivity of the diamond norm (Lemma 4) to obtain a covering number bound for the class of maps that can be implemented by a variational QMLM, understood as a parametrized CPTP map as described in the main text. Again, we first state the bound for the unitary case.

Theorem 1 (Metric entropy bounds for unitary QMLMs)**.** Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be an n-qubit QMLM with T parameterized 2-qubit unitary gates and an arbitrary number of non-trainable, global unitary gates. Let $\mathcal{U}^{\mathrm{QMLM}} \subset \mathcal{U}(\mathbb{C}^{2^n})$ denote the set of n-qubit unitaries that can be implemented by the QMLM $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$.

Then, for every $\varepsilon \in (0, 1]$, there exists an ε-covering net $\mathcal{N}_\varepsilon$ of (the set of unitary channels corresponding to) $\mathcal{U}^{\mathrm{QMLM}}$ w.r.t. the diamond distance such that the logarithm of its size can be upper bounded as

$$\log(|\mathcal{N}_\epsilon|) \leqslant 32T \log\left(\frac{12T}{\varepsilon}\right). \tag{18}$$

Proof. Let $\varepsilon \in (0, 1]$, write $\tilde{\varepsilon} := \frac{\varepsilon}{2T}$. By Lemma 6, there exists an $\tilde{\varepsilon}$-net $\tilde{\mathcal{N}}_{\tilde{\varepsilon}}$ of $\mathcal{U}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right)$ w.r.t. the spectral norm of size $|\tilde{\mathcal{N}}_{\tilde{\varepsilon}}| \leqslant (6/\tilde{\varepsilon})^{32} = (12T/\varepsilon)^{32}$.

Note that any $U \in \mathcal{U}^{\mathrm{QMLM}}$ is of the form $U = V_T U_T V_{T-1} U_{T-1} V_{T-2} \ldots V_1 U_1 V_0$, where U_t, $1 \leqslant t \leqslant T$, are a particular choice of the trainable 2-qubit unitaries and V_s, $0 \leqslant s \leqslant T+1$, are the non-trainable n-qubit unitaries occurring in the QMLM. (For ease of readability, we have not written out the tensor factors of identities accompanying the U_t.) We now consider the set of unitaries obtained by plugging the elements of $\tilde{\mathcal{N}}_{\tilde{\varepsilon}}$ as trainable 2-qubit unitaries into the QMLM. That is, we take

$$\mathcal{N}_\varepsilon := \left\{ V_T U_T V_{T-1} U_{T-1} V_{T-2} \ldots V_1 U_1 V_0 \mid U_t \in \tilde{\mathcal{N}}_{\tilde{\varepsilon}}, 1 \leqslant t \leqslant T \right\}. \tag{19}$$

Let $U \in \mathcal{U}^{\mathrm{QMLM}}$ be an arbitrary n-qubit unitary that can be implemented by the QMLM, i.e., $U = V_T U_T V_{T-1} U_{T-1} V_{T-2} \ldots V_1 U_1 V_0$ for some $U_t \in \mathcal{U}(\mathbb{C}^2 \otimes \mathbb{C}^2)$, $1 \leqslant t \leqslant T$. Let $\mathcal{U}$ denote the corresponding unitary channel. As $\tilde{\mathcal{N}}_{\tilde{\varepsilon}}$ is an $\tilde{\varepsilon}$-net for the set of 2-qubit unitaries, we can find $\tilde{U}_t \in \tilde{\mathcal{N}}_{\tilde{\varepsilon}}$, $1 \leqslant t \leqslant T$, such that $\|U_t - \tilde{U}_t\| \leqslant \tilde{\varepsilon}$ for all $1 \leqslant t \leqslant T$. Then, the unitary channel $\tilde{\mathcal{U}}$ described by $\tilde{U} := V_T \tilde{U}_T V_{T-1} \tilde{U}_{T-1} V_{T-2} \ldots V_1 \tilde{U}_1 V_0 \in \mathcal{N}_\varepsilon$ satisfies

$$\|\mathcal{U} - \tilde{\mathcal{U}}\|_\diamond \leqslant \sum_{s=0}^{T+1} \|\mathcal{V}_s - \tilde{\mathcal{V}}_s\|_\diamond + \sum_{t=1}^{T} \|\mathcal{U}_t - \tilde{\mathcal{U}}_t\|_\diamond \leqslant 2 \sum_{t=1}^{T} \|U_t - \tilde{U}_t\| \leqslant \varepsilon, \tag{20}$$

where we iteratively applied sub-additivity of the diamond distance (Lemma 4) in the first step, then used the relation between the diamond distance of unitary channels to the spectral norm distance of the corresponding unitaries (Lemma 5), and in the final step plugged in the definition of $\tilde{\varepsilon}$.

Thus, we have shown that the set of unitary channels with unitaries in $\mathcal{N}_\varepsilon$ is an ε-covering net of the set of unitary channels with unitaries in $\mathcal{U}^{\mathrm{QMLM}}$ w.r.t. the diamond distance. As $|\mathcal{N}_\varepsilon| = |\tilde{\mathcal{N}}_{\tilde{\varepsilon}}|^T$ (by definition of $\mathcal{N}_\varepsilon$), plugging in the bound on the size of $\tilde{\mathcal{N}}_{\tilde{\varepsilon}}$ then gives the desired bound on the cardinality of $\mathcal{N}_\varepsilon$ and thereby of our covering net. □

For variational quantum circuits consisting of CPTP maps, we obtain an analogous result upon replacing Lemma 6 by Lemma 7 in the previous proof:

Theorem 2 (Metric entropy bounds for QMLMs of CPTP maps)**.** Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be an n-qubit QMLM with T parameterized 2-qubit CPTP maps and an arbitrary number of non-trainable, global CPTP maps. Let $\mathcal{CPTP}^{\mathrm{QMLM}} \subset \mathcal{CPTP}\left((\mathbb{C}^2)^{\otimes n}\right)$ denote the set of n-qubit CPTP maps that can be implemented by the circuit QMLM $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$.

For any $\varepsilon \in (0,1]$, there exists an ε-covering net $\mathcal{N}_\varepsilon$ of $\mathcal{CPTP}^{\mathrm{QMLM}}$ w.r.t. the diamond distance such that the logarithm of its size can be upper bounded as

$$\log(|\mathcal{N}_\epsilon|) \leqslant 512T \log\left(\frac{6T}{\varepsilon}\right). \tag{21}$$

In both scenarios, the metric entropy grows at worst slightly super-linearly with the number of parameterized (and thus trainable) operations.

We also provide a generalization of these metric entropy bounds that is natural for the scenario in which trainable gates are reused in the quantum machine learning model:

Theorem 3 (Metric entropy bounds for QMLMs of reused CPTP maps)**.** Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be an n-qubit QMLM with T parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is used M_t times, and an arbitrary number of non-trainable, global CPTP maps. Let $\mathcal{CPTP}^{\mathrm{QMLM}} \subset \mathcal{CPTP}\left((\mathbb{C}^2)^{\otimes n}\right)$ denote the set of n-qubit CPTP maps that can be implemented by the QMLM $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$.

For any $\varepsilon \in (0,1]$, there exists an ε-covering net $\mathcal{N}_\varepsilon$ of $\mathcal{CPTP}^{\mathrm{QMLM}}$ w.r.t. the diamond distance such that the logarithm of its size can be upper bounded as

$$\log(|\mathcal{N}_\epsilon|) \leqslant 512\left(T \log\left(\frac{6T}{\varepsilon}\right) + \sum_{t=1}^{T} \log(M_t)\right). \tag{22}$$

Proof. We can use the same reasoning as in the proof of Theorems 1 and 2 to show that we can define an ε-covering net $\mathcal{N}_\varepsilon$ for $\mathcal{CPTP}^{\mathrm{QMLM}}$ (w.r.t. $\|\cdot\|_\diamond$) by plugging the elements of an $\tilde{\varepsilon}_t$-net for $\mathcal{CPTP}(\mathbb{C}^2 \otimes \mathbb{C}^2)$ into the positions of the QMLM corresponding to the t^{th} independently trainable 2-qubit map, where $\tilde{\varepsilon}_t := \frac{\varepsilon}{T \cdot M_t}$. When picking the $\tilde{\varepsilon}_t$-nets with cardinality bounded as in Lemma 7, the cardinality of $\mathcal{N}_\varepsilon$ can be bounded as

$$|\mathcal{N}_\varepsilon| \leqslant \prod_{t=1}^{T} \left(\frac{6TM_t}{\varepsilon}\right)^{512} = \left(\left(\frac{6T}{\varepsilon}\right)^{T} \cdot \left(\prod_{t=1}^{T} M_t\right)\right)^{512}. \tag{23}$$

Taking a logarithm gives the claimed metric entropy bound. □

The growth of the metric entropies in terms of T, the number of independently trainable maps, is still at most slightly super-linear. But the growth in terms of the numbers of times that the trainable maps are reused is only logarithmic.

Note that we have formulated the metric entropy bounds for the qubit case only, but they can naturally be extended to the qudit case. Then the upper bound will depend polynomially on the dimension d.

We provide one more metric entropy bound for QMLMs, which also takes the training procedure into account, in Theorem 8. Formulating this bound, however, requires us to fix some (notational) assumptions on the optimization procedure used for training. Therefore, we postpone this final metric entropy bound to Supplementary Note 3. 2. 5..

Remark 1. Both in this section and in the following ones, we formulate our results for QMLMs whose parametrized gates act on (at most) 2 qubits. Our proofs and results straightforwardly extend to the case in which the parametrized gates act on (at most) κ qubits. In particular, when going from 2- to κ-local, the T-dependence remains the same. Only the constant prefactors in the metric entropy bounds (and thus the generalization bounds) change, namely from $2 \cdot 2^4$ to $2 \cdot 2^{2\kappa}$ in the unitary case, and from $2 \cdot 2^8$ to $2 \cdot 2^{4\kappa}$ in the CPTP case. Since κ is constant, then the latter is just prefactor that does not change the scaling of our theorems.

2. Prediction error bounds for quantum machine learning models

Using well-established tools from statistical learning theory, we can derive prediction error bounds for QMLMs from the covering number bounds established in Supplementary Note 3. 1.. Before doing so, we describe our setting in detail.

During the training process, we optimize the parameters $\boldsymbol{\alpha}$ in the (CPTP map implemented by the) quantum machine learning model $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ according to some criteria and depending on the training data. Here, we write $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{k})$ if we consider both discrete, structural parameters $\boldsymbol{k}$ and continuous parameters $\boldsymbol{\theta}$. If the QMLM structure is fixed and only the continuous parameters are optimized, we write only $\boldsymbol{\theta}$ (instead of $\boldsymbol{\alpha}$). Note that we do not make any further assumptions on how the QMLM $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ depends on the parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{k})$ other than that the

discrete parameters only encode different choices of quantum circuit architectures. In particular, the dependence of the trainable gates on the continuous parameters $\boldsymbol{\theta}$ can be arbitrary.

We use an observable to quantify how good/bad the output state is, this will serve as our loss function. More concretely, for an input x_i and (classical or quantum) target output y_i, we define the loss function of the parameter setting $\boldsymbol{\alpha}$ to be

$$\ell(\boldsymbol{\alpha}; x_i, y_i) = \mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}(\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}} \otimes \mathrm{id})(\rho(x_i))\right], \tag{24}$$

for some Hermitian observables $O^{\mathrm{loss}}_{x_i,y_i}$. Here, $x \mapsto \rho(x)$ is some encoding of the classical data into quantum states that is fixed in advance.

As is common in classical learning theory, the prediction error bounds will depend on the largest (absolute) value that the loss function can attain. In our case, we therefore assume $C_{\mathrm{loss}} := \sup_{x,y} \|O^{\mathrm{loss}}_{x,y}\| < \infty$. That is, we assume that the spectral norm can be bounded uniformly over all possible loss observables.

For a training dataset $S = \{(x_i, y_i)\}_{i=1}^N$ of size $N \in \mathbb{N}$, the average loss on the training data is given by

$$\hat{R}_S(\boldsymbol{\alpha}) := \frac{1}{N}\sum_{i=1}^N \ell(\boldsymbol{\alpha}; x_i, y_i) = \frac{1}{N}\sum_{i=1}^N \mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}(\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}} \otimes \mathrm{id})(\rho(x_i))\right], \tag{25}$$

which is often referred to as the *training error* or *empirical risk*. This quantity can (in principle) be evaluated, given the parameter setting $\boldsymbol{\alpha}$ and the training data.

When we obtain a new input x, the prediction error of a parameter setting $\boldsymbol{\alpha}$ is taken to be the expected loss

$$R(\boldsymbol{\alpha}) := \mathop{\mathbb{E}}_{(x,y)\sim P}\left[\ell(\boldsymbol{\alpha}; x, y)\right] = \mathop{\mathbb{E}}_{(x,y)\sim P} \mathrm{Tr}\left[O^{\mathrm{loss}}_{x,y}(\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}} \otimes \mathrm{id})(\rho(x))\right], \tag{26}$$

where the expectation is w.r.t. the distribution P from which the training examples are generated. This quantity is called the *prediction error* or *expected risk*. The goal of any (classical or quantum) machine learning procedure is to achieve a small prediction error with high success probability.

As the underlying distribution P is usually unknown, we cannot directly evaluate the prediction error, even if we know the parameters $\boldsymbol{\alpha}$. In practice, one therefore often takes the training error as a proxy for the prediction error. However, this procedure can only succeed if the difference between the prediction and the training error, the so-called *generalization error*, is small. Our covering number bounds allow us to prove rigorous bounds on the generalization error incurred by a variational quantum machine learning method in the so-called *"Probably Approximately Correct" (PAC)* sense. That is, we provide bounds on the generalization error in terms of the desired success probability and the training data size. Thereby, our results provide guarantees on the prediction performance of a quantum machine learning model on unseen data, if that model performs well on the training data.

Our main result is the following:

Theorem 4 (Mother Theorem)**.** Let $\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}}(\cdot)$ be a QMLM with a variable structure. Suppose that, for every $k \in \mathbb{N}$, there are at most $G_\tau \in \mathbb{N}$ allowed structures with exactly τ parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is taken from a set $\mathcal{M}_t$ and used M_t times, and an arbitrary number of non-trainable, global CPTP maps. Also, for each $t \in \mathbb{N}$, let $\mathcal{E}^0_t \in \mathcal{CPTP}\left((\mathbb{C})^{\otimes 2}\right)$ be a fixed reference CPTP map. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, our optimization of the QMLM over structures and parameters w.r.t. the loss function $\ell(\boldsymbol{\alpha}; x_i, y_i) = \mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}(\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}} \otimes \mathrm{id})(\rho(x_i))\right]$ yields a (data-dependent) structure with $T = T(N)$ independently parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is taken from $\mathcal{M}_t$ and used M_t times, as well as the parameter setting $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(S)$.

Then, with probability at least $1-\delta$ over the choice of i.i.d. training data S of size N according to P,

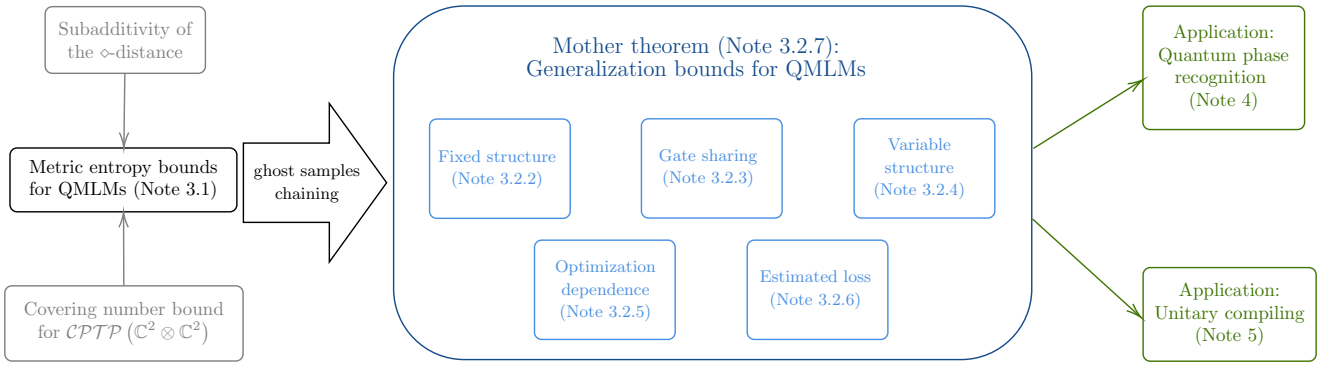
$$\begin{aligned} &R(\boldsymbol{\alpha}^*) - \hat{R}_S(\boldsymbol{\alpha}^*) \\ &\in \mathcal{O}\left(C_{\mathrm{loss}} \min\left\{\sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(M_t)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} M_t\Delta_t + \sqrt{\frac{\log(G_T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right\}\right), \end{aligned} \tag{27}$$

where $\Delta_1^T, \ldots, \Delta_T^T$ denote the (data-dependent) distance between the trainable maps in the output QMLM to the respective reference maps $\mathcal{E}^0_1, \ldots, \mathcal{E}^0_T$, $C_{\mathrm{loss}} = \sup\limits_{x,y} \|O^{\mathrm{loss}}_{x,y}\|$ is the maximum (absolute) value attainable by the loss function, and the minimum is over all $K \in \{0, \ldots, T\}$ and choices of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$.

Moreover, if the loss is not evaluated exactly, but an unbiased estimator is built from σ_{est} subsampled training data points (as in Supplementary Note 3. 2. 6.), we only incur an additional error of $\mathcal{O}\left(\sqrt{\log(1/\delta)/\sigma_{\mathrm{est}}}\right)$.

Some of the important aspects of the upper bound on the generalization error of a QMLM provided by Theorem 4 are: a dependence on the square root of the inverse of the training data size (N); an at worst slightly superlinear dependence on the number of trainable maps (T), which can improve if only a smaller number (K) of gates experience non-negligible changes during the optimization; a logarithmic dependence on the number of uses (M_t); a logarithmic dependence on the number of different architectures (G_T); and a logarithmic dependence on the reciprocal of the desired confidence level (δ).

We build up to the proof of Theorem 4 by first establishing our basic QML generalization error bound and then extending it in different directions. More precisely, we structure our presentation as follows: We start with the pedagogical Supplementary Note 3. 2. 1., in which we show a simple proof of how metric entropy bounds lead to generalization bounds, albeit not yet to their strongest form. In Supplementary Note 3. 2. 2., we demonstrate how to improve upon the simple proof strategy using a more involved technique. Then, we extend the basic generalization error bounds in multiple directions, namely to multiple copies and reused trainable maps (Supplementary Note 3. 2. 3.), variable architecture (Supplementary Note 3. 2. 4.), optimization-dependent guarantees (Supplementary Note 3. 2. 5.), and to a scenario in which we can not evaluate the loss function exactly, but only indirectly through an unbiased estimator (Supplementary Note 3. 2. 6.). Finally, we bring together all these extensions into the most general form of our result (Supplementary Note 3. 2. 7.). Our line of reasoning is summarized in Supplementary Figure 1.



Supplementary Figure 1. **Visualization of the proof structure.** We prove metric entropy bounds and use them to derive generalization bounds for different QMLM settings. We then apply our theory to quantum phase recognition and unitary compiling.

Remark 2. A loss function of the form of Eq. (24) automatically has a certain linear structure, namely, it depends linearly on the output state $(\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id})(\rho(x_i))$. Notice, however, that we can introduce also a certain type of nonlinearity through the spectral decomposition of the loss observables $O_{x_i,y_i}^{\mathrm{loss}}$. Namely, suppose that we obtain a classical output from the QMLM by measuring an observable O_{out} with spectral decomposition $O_{\mathrm{out}} = \sum_j \lambda_j |j\rangle\langle j|$. That is, given an input x_i, we output λ_j with probability $p_j(\boldsymbol{\alpha}, x_i) = \mathrm{Tr}[|j\rangle\langle j|(\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id})(\rho(x_i))] = \langle j|(\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id})(\rho(x_i))|j\rangle$. Now, we can, for example, define the loss observables as $O_{x_i,y_i}^{\mathrm{loss}} := \sum_j (y_i - \lambda_j)^2 |j\rangle\langle j|$, so that $\ell(\boldsymbol{\alpha}; x_i, y_i) = \mathbb{E}[(y_i - \lambda_j)^2]$ becomes the expected square loss between the true label y_i and our output λ_i. Here, the expectation is w.r.t. $(p_j(\boldsymbol{\alpha}, x_i))_j$. Clearly, here we can replace $(y_i - \lambda_j)^2$ by any nonlinear loss function $\tilde{\ell}(y_i, \lambda_j)$ of interest.

1. *Prelude: Metric entropy bounds imply generalization error bounds*

This section is intended to help readers not yet familiar with the theory of classical machine learning gain an intuition for how we derive our analytical results. We present a technically simple proof of a generalization bound for a fixed-architecture QMLM, which, however, is worse than that of Theorem 6 by a factor logarithmic in the training data size. Therefore, readers already well versed in statistical learning theory, or readers who want to focus on the results and not the proofs, can safely skip this pedagogical section.

We demonstrate how the metric entropy bound from Theorem 2 gives rise to a generalization bound for QMLMs with a fixed architecture, in which each trainable 2-qubit CPTP map is used only once. The simplified proof given in this section consists in combining Hoeffding's concentration inequality (Lemma 1) with a union bound over a suitable covering net. Informally, we show that it suffices to prove good generalization simultaneously for all elements in a covering net, which we can obtain from a union bound over standard concentration guarantees for each single element of the covering net.

Theorem 5 (Prediction error bound for quantum machine learning - Fixed structure (Preliminary version))**.** Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a fixed architecture consisting of T parameterized 2-qubit CPTP maps and an arbitrary number of non-trainable, global CPTP maps. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, our optimization yields the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$.

Then, with probability at least $1-\delta$ over the choice of i.i.d. training data S of size N according to P,

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(TN)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right). \tag{28}$$

Proof. For any parameter setting $\boldsymbol{\theta}$, fixed independently of the choice of training data, we see that $\ell(\boldsymbol{\theta}; x_1, y_1), \ldots, \ell(\boldsymbol{\theta}; x_N, y_N)$ are independent random variables taking values in $[-C_{\mathrm{loss}}, C_{\mathrm{loss}}]$. So Hoeffding's Lemma (Lemma 1) tells us that, $\forall \eta > 0$, have

$$\mathbb{P}_S\left[R(\boldsymbol{\theta}) - \hat{R}_S(\boldsymbol{\theta}) > \eta\right] \leqslant \exp\left(-\frac{N\eta^2}{2C_{\mathrm{loss}}^2}\right). \tag{29}$$

Here, $\mathbb{P}_S[\cdot] = \mathbb{P}_{S\sim P^N}[\cdot]$ denotes the probability over training data sets $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, with the (x_i, y_i) drawn i.i.d. from the probability measure P. Next, we let $\varepsilon = \sqrt{T/N} > 0$, take $\mathcal{N}_\varepsilon$ to be an ε-covering net of the set of CPTP maps that can be implemented by the QMLM, and we take a union bound over $\mathcal{N}_\varepsilon$, with which we obtain

$$\mathbb{P}_S\left[\exists \mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}} \in \mathcal{N}_\varepsilon :\ R(\boldsymbol{\theta}) - \hat{R}_S(\boldsymbol{\theta}) > \eta\right] \leqslant |\mathcal{N}_\varepsilon| \cdot \exp\left(-\frac{N\eta^2}{2C_{\mathrm{loss}}^2}\right). \tag{30}$$

As we took $\mathcal{N}_\varepsilon$ to be an ε-covering net (w.r.t. the diamond norm) of the class of CPTP maps that the QMLM can implement, and since $\|\mathcal{E} - \tilde{\mathcal{E}}\|_\diamond \leqslant \varepsilon$ directly implies, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$\left|\mathrm{Tr}\left[O_{x,y}^{\mathrm{loss}}(\mathcal{E}\otimes\mathrm{id})(\rho(x))\right] - \mathrm{Tr}\left[O_{x,y}^{\mathrm{loss}}(\tilde{\mathcal{E}}\otimes\mathrm{id})(\rho(x))\right]\right| \leqslant \|O_{x,y}^{\mathrm{loss}}\| \cdot \|\mathcal{E} - \tilde{\mathcal{E}}\|_\diamond \leqslant C_{\mathrm{loss}}\varepsilon, \tag{31}$$

we conclude, because of the form of the loss function ℓ, that

$$\mathbb{P}_S\left[R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) > \eta + 2C_{\mathrm{loss}}\varepsilon\right] \leqslant \mathbb{P}_S\left[\exists \mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}} \in \mathcal{N}_\varepsilon :\ \hat{R}_S(\boldsymbol{\theta}) > R(\boldsymbol{\theta}) + \eta\right] \tag{32}$$

$$\leqslant |\mathcal{N}_\varepsilon| \cdot \exp\left(-\frac{N\eta^2}{2C_{\mathrm{loss}}^2}\right) \tag{33}$$

Thus, for any $\delta \in (0,1)$, by choosing $\eta = C_{\mathrm{loss}}\sqrt{\frac{2\log(|\mathcal{N}_\varepsilon|/\delta)}{N}}$, we can guarantee that, with probability at least $1-\delta$ over the choice of training data S of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \leqslant C_{\mathrm{loss}}\sqrt{\frac{2\log(|\mathcal{N}_\varepsilon|/\delta)}{N}} + 2C_{\mathrm{loss}}\sqrt{\frac{T}{N}}. \tag{34}$$

Now, we recall that, by Theorem 2, we can take $\mathcal{N}_\varepsilon$ to satisfy $\log(|\mathcal{N}_\varepsilon|) \leqslant 512T\log(6T/\varepsilon)$. Plugging this into the previous bound, we see that, with probability at least $1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \leqslant C_{\mathrm{loss}}\sqrt{\frac{2\cdot(512T\log(6T/\varepsilon) + \log(1/\delta))}{N}} + 2C_{\mathrm{loss}}\sqrt{\frac{T}{N}} \tag{35}$$

$$\leqslant C_{\mathrm{loss}}\sqrt{\frac{2\cdot(512T\log(6\sqrt{TN}) + \log(1/\delta))}{N}} + 2C_{\mathrm{loss}}\sqrt{\frac{T}{N}} \tag{36}$$

$$\in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(TN)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right), \tag{37}$$

which is the claimed generalization error bound. □

Remark 3. At first glance, it might seem that simply plugging the parameter setting $\boldsymbol{\theta}^*$ into Eq. (29) would already give us a good concentration bound for the parameter setting $\boldsymbol{\theta}^*$ obtained through training and that the union bound over the covering net is not actually necessary in the above proof. However, as the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$ depends on the whole training data set S, the random variables $\ell(\boldsymbol{\theta}^*; x_i, y_i)$, $i = 1, \ldots, N$, are not statistically independent. Thus, Hoeffding's inequality alone cannot be used to obtain a version of Eq. (29) with $\boldsymbol{\theta}$ replaced by the data-dependent $\boldsymbol{\theta}^*$.

The generalization bound established in Theorem 5 already shows the right behavior in terms of the dependence on T, the number of trainable maps. However, the dependence on N, the sample size, still contains an undesirable logarithmic term. In classical statistical learning theory, it is well known that a proof strategy as above, based on combining Hoeffding's concentration inequality with a union bound over a covering net, incurs such a $\log(N)$-term. Fortunately, a technique for removing this term is also known and we will use it to tighten the prediction error bound in the next subsection.

2. *Basic prediction error bound for fixed architecture*

Our first prediction error bound is for the case of a variational QMLM with a fixed architecture. In particular, while the parameters in the trainable 2-qubit CPTP maps can be optimized over, the structure of the QMLM, i.e., the arrangement of the different elements, and in particular the overall depth and size, remain fixed. (We provide a generalization to variable circuit architectures in Supplementary Note 3. 2. 4..) In this scenario, we have the following generalization error bound:

Theorem 6 (Prediction error bound for quantum machine learning - Fixed structure)**.** Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a fixed architecture consisting of T parameterized 2-qubit CPTP maps and an arbitrary number of non-trainable, global CPTP maps. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, our optimization yields the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$.

Then, with probability at least $1-\delta$ over the choice of i.i.d. training data S of size N according to P,

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right). \tag{38}$$

In case the training data contains quantum labels, we assume the training data states to be reproducible so that we can use the data both for the optimization procedure and for evaluating the training error.

Proof. The proof proceeds in two steps: The first step is to upper-bound the generalization error in terms of the expected supremum of a random process. (This well known technique is described, e.g., in Theorem 3.3 in [54].) In the second step, we invoke the chaining technique to further upper-bound this expected supremum in terms of covering numbers. (This method goes back to [5]. See, e.g., Section 8 of [53] for a pedagogical presentation.) At this point, we apply our covering numbers bounds to finish the proof.

For ease of notation in the first step, we define $\varphi : (\mathcal{X}\times\mathcal{Y})^N \to \mathbb{R}$ as $\varphi(S) := \sup_{\boldsymbol{\theta}}\left\{R(\boldsymbol{\theta}) - \hat{R}_S(\boldsymbol{\theta})\right\}$, where the supremum goes over all possible parameter settings in the QMLM. We first observe that, if $S = \{(x_i, y_i)\}_{i=1}^N$ and $\tilde{S} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^N$ differ only in a single labelled example, then $|\varphi(S) - \varphi(\tilde{S})| \leqslant {}^{2C_{\mathrm{loss}}}/_N$ (because the loss function has values in $[-C_{\mathrm{loss}}, C_{\mathrm{loss}}]$). Therefore, we can apply McDiarmid's inequality (Lemma 2) and obtain that, for every $\varepsilon > 0$, $\mathbb{P}_S[\varphi(S) - \mathbb{E}_{\tilde{S}}[\varphi(\tilde{S})] \geqslant \varepsilon] \leqslant \exp\left(-{}^{N\varepsilon^2}/_{2C_{\mathrm{loss}}^2}\right)$. Hence, for every $\delta \in (0,1)$, with probability $\geqslant 1 - {}^{\delta}/_2$ over the choice of training data, we have , with $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$ as in the statement of the Theorem,

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \leqslant \varphi(S) \leqslant \mathbb{E}_{\tilde{S}}[\varphi(\tilde{S})] + C_{\mathrm{loss}}\sqrt{\frac{2\log({}^2/_\delta)}{N}}. \tag{39}$$

We now upper-bound $\mathbb{E}_{\tilde{S}}[\varphi(\tilde{S})]$. To this end, we introduce a so-called ghost sample. Namely, we take $S' = \{(x_i', y_i')\}_{i=1}^N$ to be an i.i.d. copy of $\tilde{S}$. Then, we can bound

$$\mathbb{E}_{\tilde{S}}[\varphi(\tilde{S})] = \mathbb{E}_{\tilde{S}}\left[\sup_{\boldsymbol{\theta}}\left\{\frac{1}{N}\sum_{i=1}^N\left(\mathbb{E}_{(x_i',y_i')\sim P}[\ell(\boldsymbol{\theta}; x_i', y_i')] - \ell(\boldsymbol{\theta}; \tilde{x}_i, \tilde{y}_i)\right)\right\}\right] \tag{40}$$

$$\leqslant \mathbb{E}_{\tilde{S},S'}\left[\sup_{\boldsymbol{\theta}}\left\{\frac{1}{N}\sum_{i=1}^N\left(\ell(\boldsymbol{\theta}; x_i', y_i') - \ell(\boldsymbol{\theta}; \tilde{x}_i, \tilde{y}_i)\right)\right\}\right]. \tag{41}$$

Now, we use a standard symmetrization argument with i.i.d. Rademacher random variables to further upper-bound the right hand side. That is, we let $\sigma_1, \ldots, \sigma_N$ be i.i.d. Rademacher random variables, each distributed uniformly on $\{-1, 1\}$. As multiplying $(\ell(\boldsymbol{\theta}; x_i', y_i') - \ell(\boldsymbol{\theta}; \tilde{x}_i, \tilde{y}_i))$ by -1 is equivalent to interchanging the i.i.d. copies $(\tilde{x}_i, \tilde{y}_i)$ and

(x_i', y_i'), which leaves the expectation invariant, we can introduce an additional expectation value over Rademacher variables as follows:

$$\mathbb{E}_{\tilde{S},S'}\left[\sup_{\boldsymbol{\theta}}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\ell(\boldsymbol{\theta};x_i',y_i')-\ell(\boldsymbol{\theta};\tilde{x}_i,\tilde{y}_i)\right)\right\}\right]=\mathbb{E}_{\tilde{S},S'}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\left\{\frac{1}{N}\sum_{i=1}^{N}\sigma_i\left(\ell(\boldsymbol{\theta};x_i',y_i')-\ell(\boldsymbol{\theta};\tilde{x}_i,\tilde{y}_i)\right)\right\}\right] \tag{42}$$

$$\leqslant 2\mathbb{E}_{\tilde{S}}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};\tilde{x}_i,\tilde{y}_i)\right]. \tag{43}$$

The quantity on the right hand side is not an empirical quantity, i.e., it cannot be directly evaluated only from the training data without knowledge of the underlying distribution P. However, another application of McDiarmid's inequality shows that, for every $\varepsilon>0$,

$$\mathbb{P}_S\left[\mathbb{E}_{\tilde{S}}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};\tilde{x}_i,\tilde{y}_i)\right]-\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};x_i,y_i)\right]\geqslant\varepsilon\right]\leqslant\exp\left(-\frac{N\varepsilon^2}{2C_{\text{loss}}^2}\right), \tag{44}$$

where we again used that the loss function has values in $[-C_{\text{loss}},C_{\text{loss}}]$. In other words, for every $\delta\in(0,1)$, with probability $\geqslant 1-\delta/2$ over the choice of training data, we have

$$\mathbb{E}_{\tilde{S}}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};\tilde{x}_i,\tilde{y}_i)\right]\leqslant\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};x_i,y_i)\right]+C_{\text{loss}}\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{45}$$

When applying a union bound, we can combine Eq. (39) and (45) to conclude: For every $\delta\in(0,1)$, with probability $\geqslant 1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*)-\hat{R}(\boldsymbol{\theta}^*)\leqslant 2\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};x_i,y_i)\right]+3C_{\text{loss}}\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{46}$$

This concludes the first step of the proof.

As a second step, we use chaining to upper-bound $\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};x_i,y_i)\right]$ in terms of covering numbers. For $j\in\mathbb{N}_0$, define $\alpha_j:=2^{-j}C_{\text{loss}}$. By Theorem 2, for every $j\in\mathbb{N}_0$, there exists an 2^{-j}-covering net $\mathcal{N}_j$ (w.r.t. the diamond norm) of the set of CPTP maps that can be implemented by the QMLM, satisfying $|\mathcal{N}_j|=\left(6T/2^{-j}\right)^{512T}=\left(6T\cdot 2^j\right)^{512T}$. In particular, for every $j\in\mathbb{N}$ and for every parameter setting $\boldsymbol{\theta}$, there exists a CPTP map $\mathcal{E}_{\boldsymbol{\theta},j}\in\mathcal{N}_j$ and $\|\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}-\mathcal{E}_{\boldsymbol{\theta},j}\|_{\diamond}\leqslant 2^{-j}$. For $j=0$, we can take the 1-covering net $\mathcal{N}_0=\{0\}$.

With this observation at hand, we can bound, for any $m\in\mathbb{N}$,

$$\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(\boldsymbol{\theta};x_i,y_i)\right] \tag{47}$$

$$=\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\text{Tr}\left[O_{x_i,y_i}^{\text{loss}}(\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}\otimes\text{id})(\rho(x_i))\right]\right] \tag{48}$$

$$=\frac{1}{N}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\left\{\sum_{i=1}^{N}\sigma_i\left(\text{Tr}\left[O_{x_i,y_i}^{\text{loss}}((\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}-\mathcal{E}_{\boldsymbol{\theta},m})\otimes\text{id})(\rho(x_i))\right]+\sum_{j=1}^{m}\text{Tr}\left[O_{x_i,y_i}^{\text{loss}}((\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}_{\boldsymbol{\theta},j-1})\otimes\text{id})(\rho(x_i))\right]\right)\right\}\right] \tag{49}$$

$$\leqslant\frac{1}{N}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\sigma_i\text{Tr}\left[O_{x_i,y_i}^{\text{loss}}((\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}-\mathcal{E}_{\boldsymbol{\theta},m})\otimes\text{id})(\rho(x_i))\right]\right] \tag{50}$$

$$+\frac{1}{N}\sum_{j=1}^{m}\mathbb{E}_{\sigma}\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\sigma_i\text{Tr}\left[O_{x_i,y_i}^{\text{loss}}((\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}_{\boldsymbol{\theta},j-1})\otimes\text{id})(\rho(x_i))\right]\right] \tag{51}$$

where we used the telescopic sum representation $\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}=\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}-\mathcal{E}_{\boldsymbol{\theta},m}+\sum_{j=1}^{m}(\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}_{\boldsymbol{\theta},j-1})$. We bound the two

summands appearing in Eq. (51) separately. For the first term, we can apply Hölder's inequality to obtain

$$\frac{1}{N}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\sigma_i\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}((\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}-\mathcal{E}_{\boldsymbol{\theta},m})\otimes\mathrm{id})(\rho(x_i))\right]\right] \tag{52}$$

$$\leqslant \frac{1}{N}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\left|\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}((\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}-\mathcal{E}_{\boldsymbol{\theta},m})\otimes\mathrm{id})(\rho(x_i))\right]\right|\right] \tag{53}$$

$$\leqslant \frac{1}{N}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}C_{\mathrm{loss}}\|((\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}-\mathcal{E}_{\boldsymbol{\theta},m})\otimes\mathrm{id})(\rho(x_i))\|_1\right] \tag{54}$$

$$\leqslant \frac{C_{\mathrm{loss}}}{N}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\|\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}-\mathcal{E}_{\boldsymbol{\theta},m}\|_\diamond\right] \tag{55}$$

$$\leqslant C_{\mathrm{loss}}\cdot 2^{-m} \tag{56}$$

$$= \alpha_m. \tag{57}$$

For the second term, we observe that, thanks to Minkowski's inequality, for every parameter setting $\boldsymbol{\theta}$,

$$\sqrt{\sum_{i=1}^{N}\left|\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}((\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}_{\boldsymbol{\theta},j-1})\otimes\mathrm{id})(\rho(x_i))\right]\right|^2} \tag{58}$$

$$\leqslant \sqrt{\sum_{i=1}^{N}\left|\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}((\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}})\otimes\mathrm{id})(\rho(x_i))\right]\right|^2}+\sqrt{\sum_{i=1}^{N}\left|\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}((\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}-\mathcal{E}_{\boldsymbol{\theta},j-1})\otimes\mathrm{id})(\rho(x_i))\right]\right|^2} \tag{59}$$

$$\leqslant \sqrt{N}C_{\mathrm{loss}}\left(\|\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}\|_\diamond+\|\mathcal{E}^{\mathrm{QMLM}}_{\boldsymbol{\theta}}-\mathcal{E}_{\boldsymbol{\theta},j-1}\|_\diamond\right) \tag{60}$$

$$\leqslant \sqrt{N}(\alpha_j+\alpha_{j-1}) \tag{61}$$

$$\leqslant 3\alpha_j\sqrt{N}. \tag{62}$$

Therefore, for each $1\leqslant j\leqslant m$, we can apply Massart's Lemma (Lemma 3) to the set

$$A \coloneqq \left\{\left((\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}(\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}_{\boldsymbol{\theta},j-1})(\rho(x_i))\right]\right)_{i=1}^{N}\right\}_{\mathcal{E}_{\boldsymbol{\theta},j}\in\mathcal{N}_j,\mathcal{E}_{\boldsymbol{\theta},j-1}\in\mathcal{N}_{j-1}}\subset\mathbb{R}^N \tag{63}$$

with radius $3\alpha_j\sqrt{N}$ and cardinality $\leqslant|\mathcal{N}_j|\cdot|\mathcal{N}_{j-1}|\leqslant|\mathcal{N}_j|^2$ to obtain

$$\frac{1}{N}\sum_{j=1}^{m}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}}\sum_{i=1}^{N}\sigma_i\mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i}(\mathcal{E}_{\boldsymbol{\theta},j}-\mathcal{E}_{\boldsymbol{\theta},j-1})(\rho(x_i))\right]\right]\leqslant\frac{3}{\sqrt{N}}\sum_{j=1}^{m}\alpha_j\sqrt{2\log\left(|\mathcal{N}_j|^2\right)} \tag{64}$$

$$\leqslant\frac{6}{\sqrt{N}}\sum_{j=1}^{m}\alpha_j\sqrt{512T\log\left(6T\cdot 2^j\right)}, \tag{65}$$

where we used the bound on the sizes of the covering nets in the last step.

If we now use $2^{-j}=2\int_{2^{-j-1}}^{2^{-j}}\mathrm{d}\alpha$, we can rewrite the upper bound as

$$\frac{6}{\sqrt{N}}\sum_{j=1}^{m}\alpha_j\sqrt{512T\log\left(6T\cdot 2^j\right)}=\frac{12}{\sqrt{N}}\sum_{j=1}^{m}\int_{2^{-j-1}}^{2^{-j}}C_{\mathrm{loss}}\sqrt{512T\log\left(6T\cdot 2^j\right)}\,\mathrm{d}\alpha \tag{66}$$

$$\leqslant\frac{12}{\sqrt{N}}\sum_{j=1}^{m}\int_{2^{-j-1}}^{2^{-j}}C_{\mathrm{loss}}\sqrt{512T\log\left(\frac{6T}{\alpha}\right)}\,\mathrm{d}\alpha \tag{67}$$

$$=\frac{12C_{\mathrm{loss}}}{\sqrt{N}}\int_{2^{-(m+1)}}^{2^{-1}}\sqrt{512T\log\left(\frac{6T}{\alpha}\right)}\,\mathrm{d}\alpha, \tag{68}$$

where, in the first inequality, we used that $2^j \leqslant 1/\alpha$ holds inside the limits of the integral.

Combining Eq. (57) and (68), we have proved that, for every $m \in \mathbb{N}$,

$$\mathbb{E}_\sigma \left[\sup_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i) \right] \leqslant \alpha_m + \frac{12 C_{\text{loss}}}{\sqrt{N}} \int_{2^{-(m+1)}}^{2^{-1}} \sqrt{512 T \log\left(\frac{6T}{\alpha}\right)} \, \mathrm{d}\alpha. \tag{69}$$

If we take the limit $m \to \infty$, this becomes

$$\mathbb{E}_\sigma \left[\sup_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i) \right] \leqslant \frac{12 C_{\text{loss}}}{\sqrt{N}} \sqrt{512 T} \int_0^{1/2} \sqrt{\log\left(\frac{6T}{\alpha}\right)} \, \mathrm{d}\alpha \tag{70}$$

$$\leqslant \frac{12 C_{\text{loss}}}{\sqrt{N}} \sqrt{512 T} \cdot \int_0^{1/2} \sqrt{\log(6T) + \log\left(\frac{1}{\alpha}\right)} \, \mathrm{d}\alpha \tag{71}$$

$$\leqslant \frac{12 C_{\text{loss}}}{\sqrt{N}} \sqrt{512 T} \cdot \int_0^{1/2} \sqrt{\log(6T)} + \sqrt{\log\left(\frac{1}{\alpha}\right)} \, \mathrm{d}\alpha \tag{72}$$

$$= \frac{12 C_{\text{loss}}}{\sqrt{N}} \sqrt{512 T} \cdot \left(\frac{1}{2} \sqrt{\log(6T)} + \int_0^{1/2} \sqrt{\log\left(\frac{1}{\alpha}\right)} \, \mathrm{d}\alpha \right) \tag{73}$$

$$= \frac{12 C_{\text{loss}}}{\sqrt{N}} \sqrt{512 T} \cdot \left(\frac{1}{2} \sqrt{\log(6T)} + \frac{1}{2} \sqrt{\log 2} - \frac{\sqrt{\pi}}{2} \operatorname{erf}(\sqrt{\log 2}) - \frac{\sqrt{\pi}}{2} \right), \tag{74}$$

where we used the integral $\int \sqrt{\log 1/x} \, \mathrm{d}x = x\sqrt{\log 1/x} - (\sqrt{\pi}/2) \cdot \operatorname{erf}(\sqrt{\log 1/x})$, with the error function defined as $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, \mathrm{d}t$.

We can now combine Eq. (46) with (74) and obtain: With probability $\geqslant 1 - \delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \leqslant \frac{24 C_{\text{loss}}}{\sqrt{N}} \sqrt{512 T} \cdot \left(\frac{1}{2} \sqrt{\log(6T)} + \frac{1}{2} \sqrt{\log 2} - \frac{\sqrt{\pi}}{2} \operatorname{erf}(\sqrt{\log 2}) - \frac{\sqrt{\pi}}{2} \right) + 3 C_{\text{loss}} \sqrt{\frac{2 \log(2/\delta)}{N}} \tag{75}$$

$$\in \mathcal{O}\left(C_{\text{loss}} \left(\sqrt{\frac{T \log(T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} \right) \right), \tag{76}$$

which is the claimed prediction error bound. □

Remark 4. For simplicity, throughout the proof of Theorem 6 we have treated $\boldsymbol{\theta}^*(S)$ as a deterministic function of S. However, the proof extends to the case in which the parameter setting $\boldsymbol{\theta}^*(S)$ is a random variable depending on S. Then, the generalization error bound would hold with high probability over the choice of the training data and over the internal randomness of the optimization procedure. This is the case for all our prediction error bounds and is important because quantum subroutines in QML procedures make them inherently probabilistic.

Remark 5. A conceptual difference between the proof of Theorem 5 and that of Theorem 6, which can also be seen as an underlying reason for why the latter leads to a tighter bound than the former, is the following: To prove Theorem 5, we used a $\sqrt{T/N}$-covering net for the set of CPTP maps that the QMLM can implement. This can be seen as measuring the complexity of the QMLM at a single specific resolution, namely the resolution $\varepsilon = \sqrt{T/N}$. In contrast, the proof of Theorem 6 considers a complexity measure for the QMLM obtained by averaging over complexities (here, covering numbers) at multiple different resolutions. Thus, from a high-level view, the chaining-based proof strategy for Theorem 6 improves upon the reasoning behind Theorem 5 by taking multiple resolutions into account.

Theorem 6 can be interpreted as follows: By taking the training data size N to effectively scale linearly in the number of trainable elements T, we can ensure that a small training error also implies a small prediction error (with high probability).

In the following, we describe extensions of Theorem 6 to different scenarios of interest, and then summarize these in a general "mother theorem."

3. Extension to multiple copies and gate-sharing

In practice, one often employs quantum machine learning models that reuse the same parameterized gates multiple times, such as quantum convolutional neural networks (QCNNs) [25]. In such a scenario, we speak of "gate-sharing". While the number of trainable elements in such models can still be large, only few of them can be trained independently. As a first extension of Theorem 6, we show that the generalization performance of such a models is determined by the effective number of independently trainable elements.

Corollary 1. Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a fixed architecture consisting of T independently parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is used M_t times, and an arbitrary number of non-trainable, global CPTP maps. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, our optimization yields the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$.

Then, with probability at least $1-\delta$ over the choice of i.i.d. training data S of size N according to P,

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{\sum_{t=1}^T \log(M_t)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right). \tag{77}$$

Proof. The proof strategy is the same as for Theorem 6, we only change the covering number bound to be applied. Namely, instead of Theorem 2, we use Theorem 3.

More precisely, we recall Eq. (46), which tells us: For every $\delta \in (0,1)$, with probability $\geqslant 1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \leqslant 2\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^N \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i)\right] + 3C_{\mathrm{loss}}\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{78}$$

And, with the same chaining technique as detailed in the proof of Theorem 6, we can bound the above expectation over Rademacher random variables as

$$\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^N \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i)\right] \leqslant \frac{24C_{\mathrm{loss}}}{\sqrt{N}} \int_0^{1/2} \sqrt{\log\left(\mathcal{N}(\mathcal{CPTP}^{\mathrm{QMLM}}, \|\cdot\|_\diamond, \alpha)\right)}\, \mathrm{d}\alpha, \tag{79}$$

where we used the notation from Theorem 3 for the set $\mathcal{CPTP}^{\mathrm{QMLM}}$ of n-qubit CPTP maps that can be implemented by the QMLM. Now, we use the metric entropy bound proved in Theorem 3 to further upper bound the integral as

$$\int_0^{1/2} \sqrt{\log\left(\mathcal{N}(\mathcal{CPTP}_{\mathcal{A}}, \|\cdot\|_\diamond, \alpha)\right)}\, \mathrm{d}\alpha \leqslant \int_0^{1/2} \sqrt{512\left(T\log\left(\frac{6T}{\alpha}\right) + \sum_{t=1}^T \log(M_t)\right)}\, \mathrm{d}\alpha \tag{80}$$

$$\leqslant \sqrt{512T\log(6T)} \int_0^{1/2} \sqrt{\log\left(\frac{1}{\alpha}\right)}\, \mathrm{d}\alpha + \frac{\sqrt{512}}{2}\sqrt{\sum_{t=1}^T \log(M_t)}. \tag{81}$$

As $x \mapsto \log(1/x)$ has an integrable singularity at $x = 0$, the integral in this expression is simply a multiplicative constant. Therefore, after plugging in the bound of Eq. (81) into Eq. (79), and then plugging the resulting bound on the Rademacher expectation into Eq. (78), we obtain: For every $\delta \in (0,1)$, with probability $\geqslant 1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{\sum_{t=1}^T \log(M_t)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right), \tag{82}$$

the claimed generalization bound. □

A naive approach to the scenario of Corollary 1 would be to upper-bound the metric entropy, and thus the prediction error, in terms of the total number of trainable elements in the QMLM. That, however, would lead to a significantly worse dependence of the prediction error bound on M_t, the numbers of uses, namely, of the form

$$C_{\mathrm{loss}}\sqrt{\frac{T\left(\sum_{t=1}^T M_t\right)\log\left(T\sum_{t=1}^T M_t\right)}{N}}. \tag{83}$$

Our more careful analysis shows the tighter bound in which the numbers of uses M_t only appear logarithmically, which is crucial for our application of the bound to quantum phase recognition with QCNNs (see Supplementary Note 4.). This is possible because, even though there are in principle $T\sum_{t=1}^{T} M_t$ trainable elements in the quantum neural network, they are not trained independently. Rather, the parameter setting for the t^{th} parameterized 2-qubit CPTP map is reused M_t times. This clearly shows that reusing parameters is, from a generalization perspective, preferable to having more independent parameters.

As a special case of Corollary 1, we obtain a prediction error bound for the scenario in which multiple copies of a QMLM (with the same parameter settings) are run in parallel:

Corollary 2. Let $\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}}(\cdot)$ be a QMLM with a fixed architecture consisting of T independently parameterized 2-qubit CPTP maps and an arbitrary number of non-trainable, global CPTP maps. By using M copies of this model in parallel, we can consider loss functions of the form

$$\ell(\boldsymbol{\theta}; x, y) = \operatorname{Tr}\left[O_{x,y}^{\text{loss}}\left((\mathcal{E}_{\boldsymbol{\theta}}^{\text{QMLM}} \otimes \mathrm{id})(\rho(x))\right)^{\otimes M}\right], \tag{84}$$

where $O_{x,y}^{\text{loss}}$ are observables acting on the M-fold tensor product of an n-qubit system. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^{N}$ of size N, our optimization yields the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$.

Then, with probability at least $1 - \delta$ over the choice of i.i.d. training data S of size N according to P,

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{T\log(M)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right). \tag{85}$$

Once we observe that $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b} \leqslant \sqrt{2(a+b)}$ holds for all $a, b \geqslant 0$, we see that the upper bound in Corollary 2 can be rewritten as

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}}\left(\sqrt{\frac{T\log(TM)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right). \tag{86}$$

If δ is taken to be a fixed desired accuracy level and C_{loss} is also considered to be a fixed constant, this becomes the bound stated in Theorem 2 in the main text.

Corollary 2 tells us that, even when using many copies of the QMLM, as the expressiveness of the corresponding function class grows at most logarithmically with the number of copies, we can still obtain a good prediction error. Note that, as in Corollary 1, it is crucial that the same parameter setting is used for each copy.

We can also phrase the result as follows: We can upper-bound the prediction error incurred when using multiple copies of a quantum machine learning model for evaluating the loss by an expression that depends crucially on the number of trainable elements per copy, but only mildly on the number of copies.

Remark 6. Cases of interest that Corollary 2 describes are, e.g., the loss functions obtained by first performing (independent) product measurements on each of the M copies, then taking an average (for a continuous target space) or a majority vote (for a discrete target space) of the obtained measurement outcomes, and finally post-processing this value by a classical loss function (such as the squared error loss). Such procedures arise naturally when taking into account that multiple shots are needed to accurately estimate the expectation value of an observable measured on the QMLM output. Note, however, that we cannot apply arbitrary procedures for post-processing single-copy measurement outcomes and still hope for a good prediction error. If C_{loss}, which here is the supremum over the spectral norms of the M-copy observables $O_{x,y}^{\text{loss}}$, scales badly (e.g., linearly) with M, the prediction error bound does so as well.

4. Extension to variable circuit architecture

For practical purposes, it might not be advantageous to fix the number of trainable elements in the QMLM, or even its structure more generally, in advance. Rather, one might also want to optimize over a discrete set of possible architectures, e.g., by growing or truncating the QMLM during the training phase. Therefore, in our second extension of Theorem 6, we provide a prediction error bound for such a variable structure scenario.

Corollary 3. Let $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a variable structure. Suppose that, for every $\tau \in \mathbb{N}$, there are at most $G_\tau \in \mathbb{N}$ allowed structures with exactly τ parameterized 2-qubit CPTP maps and an arbitrary number of non-trainable, global CPTP maps. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, our optimization yields a (data-dependent) structure with $T = T(S)$ parameterized 2-qubit CPTP maps and the parameter setting $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(S)$.

Then, with probability at least $1 - \delta$ over the choice of i.i.d. training data S of size N according to P,

$$R(\boldsymbol{\alpha}^*) - \hat{R}_S(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{\log G_T}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right). \tag{87}$$

Proof. By Theorem 6, for every $\tau \in \mathbb{N}$, for every one of the G_τ allowed structures with exactly τ parameterized 2-qubit CPTP maps, with probability $\geqslant 1 - \delta/2G_\tau\tau^2$ over the choice of i.i.d. training data S of size N according to P, if $\boldsymbol{\theta}_\tau^*$ is a (continuous) parameter setting (for the τ parameterized maps) obtained through optimization upon input of training data S, we have the generalization error bound

$$\mathop{\mathbb{E}}_{(x,y)\sim P}[\ell(\boldsymbol{\theta}_\tau^*; x, y)] - \frac{1}{N}\sum_{i=1}^N \ell(\boldsymbol{\theta}_\tau^*; x_i, y_i) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{\tau\log(\tau)}{N}} + \sqrt{\frac{\log(2G_\tau\tau^2/\delta)}{N}}\right)\right). \tag{88}$$

Thus, first taking a union bound over the G_τ structures with exactly τ parameterized 2-qubit CPTP maps, and then a union bound over $\tau \in \mathbb{N}$, we see: With probability $\geqslant 1 - \sum_\tau \delta/2\tau^2 \geqslant 1 - \delta$ over the choice of i.i.d. training data S of size N according to P, if the optimization upon input of data S outputs a QMLM architecture with $T = T(N)$ parameterized 2-qubit CPTP maps and the (continuous and discrete) parameter setting $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(S)$

$$R(\boldsymbol{\alpha}^*) - \hat{R}_S(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{\log(2G_T T^2/\delta)}{N}}\right)\right) \tag{89}$$

$$\in \mathcal{O}\left(C_{\mathrm{loss}}\left(\sqrt{\frac{T\log(T)}{N}} + \sqrt{\frac{\log G_T}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right), \tag{90}$$

as claimed. □

We can understand Corollary 3 as saying that the prediction error of a QMLM with a variable structure depends strongly (namely linearly) on T, the number of trainable elements that is used in the output structure of the optimization procedure, but only mildly (namely logarithmically) on G_T, the number of different possible structures with the same number of gates as the output structure. Note that the bound does not depend on all structures potentially considered during the optimization, but only on a subset of those. In particular, if the number T of trainable 2-qubit maps is fixed in advance, optimizing not only over the parameter settings of the model but also over exponentially-in-T many structures with T trainable elements does not worsen the asymptotic behavior of the generalization error.

5. *Extension taking the optimization into account*

In our previous results, we have provided bounds on the generalization error that depended on the QMLM, e.g., via the number of trainable elements or the number of copies, or even on how many different architectures are admissible. So far, however, the bounds are agnostic w.r.t. the procedure used to train the QMLM. In this section, we refine our approach to prove optimization-dependent generalization bounds, that explicitly take properties of the training process into account.

Take $\mathcal{M}_1, \ldots, \mathcal{M}_T$ to be sets of 2-qubit CPTP maps. Each set $\mathcal{M}_t$ denotes the space of 2-qubit CPTP maps that one permits for the t^{th} trainable map during the training of the QMLM $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}$. Hence, each $\mathcal{M}_t$ should be seen as the trainable space for a particular gate in the QMLM. For example, $\mathcal{M}_t$ could be the space of all 2-qubit unitary channels, or the space of all tensor products of single-qubit CPTP maps.

As discussed in the proofs of Lemmas 6 and 7, as $\mathcal{CPTP}(\mathbb{C}^2 \otimes \mathbb{C}^2)$ is compact, for every $1 \leqslant t \leqslant T$, there exists a constant $c_t \geqslant 1$, depending, e.g., on the diameter and on the effective ambient dimension of $\mathcal{M}_t$, such that

$$\log(\mathcal{N}(\mathcal{M}_t, \|\cdot\|_\diamond, \epsilon)) \leqslant c_t \log\left(1 + \frac{1}{\epsilon}\right). \tag{91}$$

Note that, as a worst-case estimate, we have $c_t \leqslant 1024$. This can be seen by arguing as in the proofs of Lemmas 6 and 7, with ambient dimension 512 and diameter 2, and then applying Bernoulli's inequality.

Given a fixed choice of parameters $\boldsymbol{\theta}$, and thereby a fixed choice $\mathcal{E}_1 \in \mathcal{M}_1, \ldots, \mathcal{E}_T \in \mathcal{M}_T$ of the trainable 2-qubit CPTP maps, the (fixed-architecture) QMLM implements the n-qubit CPTP map

$$\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}} = \mathcal{E}^{\mathrm{QMLM}}(\mathcal{E}_1, \ldots, \mathcal{E}_T) := \left(\prod_{t=1}^{T} \mathcal{F}_t \mathcal{E}_t \right) \mathcal{F}_0, \tag{92}$$

where the $\mathcal{F}_t$, for $0 \leqslant t \leqslant T$, are fixed (potentially global) CPTP maps.

Suppose that we begin the optimization of the QMLM from an initial point independent of the training data $S = \{(x_i, y_i)\}_{i=1}^N$, described by a parameter setting $\boldsymbol{\theta}_0$. We denote the choices for the T trainable maps appearing in the initial CPTP map by

$$\mathcal{E}_1^0 \in \mathcal{M}_1, \ldots, \mathcal{E}_T^0 \in \mathcal{M}_T. \tag{93}$$

After utilizing the training data for multiple rounds of optimization, the training of the QMLM finishes at a (data-dependent) point in $\mathcal{CPTP}\left((\mathbb{C})^{\otimes n}\right)$, described by a parameter vector $\boldsymbol{\theta}^*$, which we denote by the choice

$$\mathcal{E}_1^* \in \mathcal{M}_1, \ldots, \mathcal{E}_T^* \in \mathcal{M}_T, \tag{94}$$

of trainable maps. Note that $\mathcal{E}_1^*, \ldots, \mathcal{E}_T^*$ depend on the training data S. For each of the T trainable local CPTP maps M_t, we denote the distance (measured w.r.t. $\|\cdot\|_\diamond$ between the initial and the final point of the training procedure by

$$\Delta_t = \|\mathcal{E}_t^* - \mathcal{E}_t^0\|_\diamond \leqslant 2, \text{ for } t = 1, \ldots, T. \tag{95}$$

In the following Theorem, we provide a generalization guarantee for the resulting QMLM defined by the choice of trainable local maps $\mathcal{E}_1^*, \ldots, \mathcal{E}_T^*$ in terms of the optimization distances Δ_t, the number T of trainable maps, and the number N of training data points.

Theorem 7 (Optimization-dependent prediction error bound for quantum machine learning)**.** Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a fixed architecture consisting of T parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is taken from $\mathcal{M}_t$, and an arbitrary number of non-trainable, global CPTP maps. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^N$ of size N, the optimization procedure yields the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$. As described above, denote by $\Delta_t = \Delta_t(S)$ the optimization distance (measured in diamond norm) of the t^{th} trainable map.

Then, with probability $\geqslant 1 - \delta$ over the choice of i.i.d. training data S of size $N \geqslant 4$ according to P,

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}} \min\left\{ \sqrt{\frac{K \max\limits_{1 \leqslant t \leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K \log(T)}{N}} + \sum_{\substack{t=1 \\ t \neq t_1, \ldots, t_K}}^{T} \Delta_t + \sqrt{\frac{\log(1/\delta)}{N}} \right\} \right), \tag{96}$$

where the minimum is over all $K \in \{0, \ldots, T\}$ and choices of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$.

The proof of Theorem 7 again hinges on a metric entropy bound, this time for the class of CPTP maps that can be reached by the QMLM under the optimization procedure. Hence, let us first prove the following theorem.

Theorem 8. Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a fixed architecture consisting of T parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is taken from $\mathcal{M}_t$, and an arbitrary number of non-trainable, global CPTP maps. Let $0 \leqslant \Delta_1, \ldots, \Delta_T \leqslant 2$ be a sequence of distances for the trainable 2-qubit CPTP maps. Let $\mathcal{CPTP}_{(\Delta_t)_t}^{\mathrm{QMLM}} \subset \mathcal{CPTP}\left((\mathbb{C}^2)^{\otimes n}\right)$ denote the set of n-qubit CPTP maps that can be implemented by the QMLM, under the additional restriction that the t^{th} trainable gate is at most diamond-distance Δ_t away from the fixed initial point $\mathcal{E}_t^0$.

Let $K \in \{0, \ldots, T\}$. Let $t_1, \ldots, t_K \in \{1, \ldots, T\}$ be pairwise distinct. Then, for any $\varepsilon \in (0, 1]$, if we write

$$\varepsilon_K := \varepsilon + \sum_{\substack{t=1 \\ t \neq t_1, \ldots, t_K}}^{T} \Delta_t \tag{97}$$

there exists an ε_K-covering net $\mathcal{N}_{\varepsilon_K}$ of $\mathcal{CPTP}_{(\Delta_t)_t}^{\mathrm{QMLM}}$ w.r.t. the diamond distance such that the logarithm of its size can be upper bounded as

$$\log(|\mathcal{N}_{\varepsilon_K}|) \leqslant K \log(T) + K \max_{1 \leqslant t \leqslant T} c_t \log\left(1 + \frac{K}{\varepsilon}\right). \tag{98}$$

Proof. By the assumptions on the structure of $\mathcal{CPTP}^{\mathrm{QMLM}}_{(\Delta_t)_t}$, there exist fixed, global CPTP maps $\mathcal{F}_0, \ldots, \mathcal{F}_T \in \mathcal{CPTP}\left((\mathbb{C}^2)^{\otimes n}\right)$ such that any $\mathcal{E} \in \mathcal{CPTP}^{\mathrm{QMLM}}_{(\Delta_t)_t}$ can be written as $\mathcal{E} = \mathcal{F}_T\mathcal{E}_T\mathcal{F}_{T-1}\ldots\mathcal{F}_1\mathcal{E}_1\mathcal{F}_0$ for some 2-qubit CPTP maps $\mathcal{E}_t \in \mathcal{M}_t$, $1 \leqslant t \leqslant T$ such that $\|\mathcal{E}_t - \mathcal{E}_t^0\|_\diamond \leqslant \Delta_t$.

As discussed above, for each $1 \leqslant t \leqslant T$, we can take $\mathcal{N}_t$ to be an (ε/K)-covering net for $\mathcal{M}_t$ w.r.t. $\|\cdot\|_\diamond$ whose cardinality satisfies $\log(|\mathcal{N}_t|) \leqslant c_t \log(1 + K/\epsilon)$. We define $\mathcal{N}_{\varepsilon_K}$ to be the set of CPTP maps that can be implemented by $\mathcal{A}$ if exactly K of the trainable 2-qubit CPTP maps are taken from $\mathcal{N}_{t_1}, \ldots, \mathcal{N}_{t_K}$, respectively, and the last $T-K$ trainable maps are left at the initial point of the optimization. That is, we define

$$\mathcal{N}_{\varepsilon_K} := \left\{ \mathcal{E}(\tilde{\mathcal{E}}_1, \ldots, \tilde{\mathcal{E}}_T) = \left(\prod_{t=1}^{T} \mathcal{F}_t\tilde{\mathcal{E}}_t\right)\mathcal{F}_0 \;\middle|\; |\{1 \leqslant t \leqslant T \mid \tilde{\mathcal{E}}_t \neq M_t^0\}| = K \text{ and } \tilde{\mathcal{E}}_t \in \mathcal{N}_t \text{ whenever } \tilde{\mathcal{E}}_t \neq \mathcal{E}_t^0 \right\}. \tag{99}$$

Using the subadditivity of the distance induced by the diamond norm (Lemma 4), it is easy to see that, for every $\mathcal{E} = \mathcal{E}(\mathcal{E}_1, \ldots, \mathcal{E}_T) \in \mathcal{CPTP}^{\mathrm{QMLM}}_{(\Delta_t)_t}$, there exists an $\tilde{\mathcal{E}} = \mathcal{E}(\tilde{\mathcal{E}}_1, \ldots, \tilde{\mathcal{E}}_T) \in \mathcal{N}_{\varepsilon_K}$ s.t.

$$\|\mathcal{E} - \tilde{\mathcal{E}}\|_\diamond \leqslant K \cdot \frac{\varepsilon}{K} + \sum_{\substack{t=1 \\ t \neq t_1, \ldots, t_K}}^{T} \Delta_t = \varepsilon_K. \tag{100}$$

Thus, $\mathcal{N}_{\varepsilon_K}$ is indeed an ε_K-covering net for $\mathcal{CPTP}^{\mathrm{QMLM}}_{(\Delta_t)_t}$, as claimed.

It remains to observe that, by definition of $\mathcal{N}_{\varepsilon_K}$, we have

$$\begin{aligned}
\log(|\mathcal{N}_{\varepsilon_K}|) &= \log\left(\binom{T}{K} \cdot \prod_{k=1}^{K} |\mathcal{N}_t|\right) && (101)\\
&\leqslant K\log(T) + \left(\sum_{k=1}^{K} c_t\right)\log\left(1 + \frac{K}{\epsilon}\right) && (102)\\
&\leqslant K\log(T) + K \max_{1 \leqslant t \leqslant T} c_t \log\left(1 + \frac{K}{\varepsilon}\right), && (103)
\end{aligned}$$

as claimed. □

Armed with this metric entropy bound, we can now prove Theorem 7.

Proof of Theorem 7. Starting from the metric entropy bound of Theorem 8, we again argue as in the proof of Theorem 6. Recall that the first step of said proof was to establish Eq. (46). This was then followed in a second step by upper-bounding the obtained expression using a covering number integral. The first step, leading to Eq. (46), is also valid in the scenario of this Theorem. That is, we again have that, with probability $\geqslant 1-\delta$ over the choice of training data of size N,

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \leqslant 2\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i)\right] + 3C_{\mathrm{loss}}\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{104}$$

However, we have to change the second step. To this end, we first observe that, by a reasoning completely analogous the one leading up to Eq. (69), for every $m \in \mathbb{N}_0$,

$$\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i)\right] \leqslant C_{\mathrm{loss}} \cdot 2^{-m} + \frac{12C_{\mathrm{loss}}}{\sqrt{N}} \int_{2^{-(m+1)}}^{2^{-1}} \sqrt{\log\left(\mathcal{N}(\mathcal{CPTP}^{\mathrm{QMLM}}_{(\Delta_t)_t}, \|\cdot\|_\diamond, \alpha\right)}\, \mathrm{d}\alpha, \tag{105}$$

where we used the notation from Theorem 8. Fix a $K \in \{0, \ldots, T\}$ and pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$ such that

$$\tilde{\Delta} := \sum_{\substack{t=1 \\ t \neq t_1, \ldots, t_K}}^{T} \Delta_t < \frac{1}{2} \tag{106}$$

and take $m \in \mathbb{N}_0$ such that $\tilde{\Delta} < 2^{-(m+1)} < 2\tilde{\Delta}$. Then in particular $2^{-m} \leqslant 4\tilde{\Delta}$ and we can further upper bound the expression in Eq. (105) as

$$\mathbb{E}_\sigma \left[\sup_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \ell(\boldsymbol{\theta}; x_i, y_i) \right] \tag{107}$$

$$\leqslant 4C_{\text{loss}}\tilde{\Delta} + \frac{12C_{\text{loss}}}{\sqrt{N}} \int_{\tilde{\Delta}}^{2^{-1}} \sqrt{\log\left(\mathcal{N}(\mathcal{CPTP}^{\text{QMLM}}_{(\Delta_t)_t}, \|\cdot\|_\diamond, \alpha\right)} \, \mathrm{d}\alpha \tag{108}$$

$$= 4C_{\text{loss}}\tilde{\Delta} + \frac{12C_{\text{loss}}}{\sqrt{N}} \int_{0}^{2^{-1}-\tilde{\Delta}} \sqrt{\log\left(\mathcal{N}(\mathcal{CPTP}^{\text{QMLM}}_{(\Delta_t)_t}, \|\cdot\|_\diamond, \alpha + \tilde{\Delta}\right)} \, \mathrm{d}\alpha. \tag{109}$$

At this point, we can apply the metric entropy bound from Theorem 8 to obtain

$$\int_{0}^{2^{-1}-\tilde{\Delta}} \sqrt{\log\left(\mathcal{N}(\mathcal{CPTP}^{\text{QMLM}}_{(\Delta_t)_t}, \|\cdot\|_\diamond, \alpha + \tilde{\Delta}\right)} \, \mathrm{d}\alpha \leqslant \int_{0}^{2^{-1}-\tilde{\Delta}} \sqrt{K \max_{1\leqslant t\leqslant T} c_t \log\left(1 + \frac{K}{\alpha}\right)} \, \mathrm{d}\alpha \tag{110}$$

$$\leqslant \sqrt{K \max_{1\leqslant t\leqslant T} c_t} \int_{0}^{2^{-1}-\tilde{\Delta}} \sqrt{\log\left(\frac{2K}{\alpha}\right)} \, \mathrm{d}\alpha \tag{111}$$

$$\leqslant \mathcal{O}\left(\sqrt{K \max_{1\leqslant t\leqslant T} c_t \log(K)}\right). \tag{112}$$

Altogether, so far we have shown that, for any fixed choice of $K \in \{0, \ldots, T\}$ and of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$ such that $\sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t < \frac{1}{2}$, with probability $\geqslant 1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}}\left(\sqrt{\frac{K \max_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t + \sqrt{\frac{2\log(1/\delta)}{N}}\right)\right). \tag{113}$$

After a union bound over the at most $\binom{T}{K} \leqslant T^K$ different choices of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$ (with $\sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t < \frac{1}{2}$), we see that, with probability $\geqslant 1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}}\left(\sqrt{\frac{K \max_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t + \sqrt{\frac{2\log(T^K/\delta)}{N}}\right)\right) \tag{114}$$

$$\in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t + \sqrt{\frac{2\log(1/\delta)}{N}}\right\}\right), \tag{115}$$

where $K \in \{0, \ldots, T\}$ is still fixed and the minimum is over all choices of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$.

Finally, we can take a union bound over at most $T+1$ different values of K and obtain that, with probability $\geqslant 1-\delta$ over the choice of training data of size N, we have

$$R(\boldsymbol{\theta}^*) - \hat{R}(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t + \sqrt{\frac{2\log(T/\delta)}{N}}\right\}\right) \tag{116}$$

$$\in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} \Delta_t + \sqrt{\frac{2\log(1/\delta)}{N}}\right\}\right), \tag{117}$$

where the minimum is over all values of K and over all choices of pairwise distinct $t_1, \dots, t_K \in \{1, \dots, T\}$. □

If, in the generalization error bound of Theorem 7, we disregard the potential improvements gained from the $\mathcal{M}_t$-dependent constants c_t and instead replace all of them by their worst-case value 1024, we can simplify the bound to

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K\log(T)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\dots,t_K}}^{T} \Delta_t + \sqrt{\frac{\log(1/\delta)}{N}}\right\}\right), \tag{118}$$

because $K\log(K) \leqslant K\log(T)$ for all $K = 1, \dots, T$. Moreover, instead of taking a minimum over all such K and over all choices of pairwise distinct $t_1, \dots, t_K \in \{1, \dots, T\}$, we can take the minimum only over K, and fix the choice $t_k = k$ to obtain

$$R(\boldsymbol{\theta}^*) - \hat{R}_S(\boldsymbol{\theta}^*) \in \mathcal{O}\left(C_{\text{loss}} \min_{K=1,\dots,T}\left\{\sqrt{\frac{K\log(T)}{N}} + \sum_{t=K+1}^{T} \Delta_t + \sqrt{\frac{\log(1/\delta)}{N}}\right\}\right). \tag{119}$$

If we again fix a confidence level δ and consider C_{loss} as a fixed constant of the problem, this becomes the bound given in Theorem 3 for the case $M = 1$. (The case for general M follows from our "mother theorem," see Supplementary Note 3. 2. 7..)

We can clearly see that, if the optimization has only made substantial changes to few trainable maps, then the generalization error bound in Theorem 7 is dominated by the maps that have undergone more significant changes during optimization. The number of such parameterized maps could be much smaller than the overall number of the trainable CPTP maps T. Therefore, this optimization-dependent generalization error bound can significantly outperform the previous bounds, which did not take the optimization procedure into account, w.r.t. the dependence on T.

One consequence of this theorem is that a good choice of initialization for the optimization of a QMLM can not only serve to improve the cost of the optimization itself, but it can also help the generalization behavior. Namely, a particularly good choice of initialization, potentially found through pretraining on an independent data set, can lead to an optimization procedure that does not have to deviate too far from the initialization w.r.t. some of the trainable maps, which, according to our bound, will be advantageous for generalization.

A second implication of this result for what to take into account in designing an optimization procedure for training a QMLM is the following: Making large steps only on few trainable gates and only negligibly small steps on the remaining ones is, from a generalization perspective, preferable to making steps of comparable, non-negligible sizes on many (or even all) of the trainable gates.

Remark 7. We note that in the proof of Theorem 7, it was not necessary that the fixed CPTP maps $\mathcal{E}_1^0 \in \mathcal{M}_1, \dots, \mathcal{E}_T^0 \in \mathcal{M}_T$ were given as the initialization of the optimization procedure. In fact, we can take these maps to be any fixed "reference points" w.r.t. which we measure distances. Th proof then works without changes, as long as the reference maps are indeed fixed in advance, independently of the training data.

6. *Extension to unbiased estimates of measurement statistics*

In practice, we cannot obtain the exact value of $\operatorname{Tr}\left[O^{\text{loss}}_{x_i,y_i}(\mathcal{E}^{\text{QMLM}}_{\boldsymbol{\theta}} \otimes \text{id})(\rho(x_i))\right]$ for a training example (x_i, y_i) if we only perform finitely many measurements. Instead, as a proxy for the training error, we consider an unbiased estimator: For $1 \leqslant \sigma \leqslant \sigma_{\text{est}}$, with $\sigma_{\text{est}} \in \mathbb{N}$ fixed, we independently pick i_σ uniformly at random from $\{1, \dots, N\}$ and measure the observable $O^{\text{loss}}_{x_{i_\sigma},y_{i_\sigma}}$ on the output state $\mathcal{E}^{\text{QMLM}}_{\boldsymbol{\theta}}(\rho(x_{i_\sigma}))$ to yield a single measurement outcome $o^{\text{loss}}_{\boldsymbol{\theta},\sigma} \in \left[-\|O^{\text{loss}}_{x_{i_\sigma},y_{i_\sigma}}\|, \|O^{\text{loss}}_{x_{i_\sigma},y_{i_\sigma}}\|\right]$. As $\mathbb{E}\left[o^{\text{loss}}_{\boldsymbol{\theta},\sigma}\right] = \frac{1}{N}\sum_{i=1}^{N} \operatorname{Tr}\left[O^{\text{loss}}_{x_i,y_i}(\mathcal{E}^{\text{QMLM}}_{\boldsymbol{\theta}} \otimes \text{id})(\rho(x_i))\right]$, where the expectation is taken w.r.t. the sampling of i_σ and the randomness in obtaining the measurement outcome. This yields a finite sequence of observations

$$o^{\text{loss}}_{\boldsymbol{\theta},1}, \dots, o^{\text{loss}}_{\boldsymbol{\theta},\sigma_{\text{est}}}, \quad \text{with} \quad \frac{1}{\sigma_{\text{est}}}\sum_{\sigma=1}^{\sigma_{\text{est}}} o^{\text{loss}}_{\boldsymbol{\theta},\sigma} \approx \frac{1}{N}\sum_{i=1}^{N} \operatorname{Tr}\left[O^{\text{loss}}_{x_i,y_i}(\mathcal{E}^{\text{QMLM}}_{\boldsymbol{\theta}} \otimes \text{id})(\rho(x_i))\right]. \tag{120}$$

In this scenario, where we only obtain a noisy estimate of the training error from σ_{est} measurements, the prediction performance guarantee takes the following form:

Corollary 4. Let $\mathcal{E}_{\boldsymbol{\theta}}^{\mathrm{QMLM}}(\cdot)$ be a quantum machine learning model with a fixed architecture consisting of T parameterized 2-qubit CPTP maps. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^{N}$ of size N, our optimization yields the parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$.

Then, with probability at least $1 - \delta$ over the choice of i.i.d. training data S of size N according to P, over the sampling of $i_1, \ldots, i_{\sigma_{\mathrm{est}}}$, and over the σ_{est} obtained measurement outcomes,

$$\mathop{\mathbb{E}}_{x,y} \ell(\boldsymbol{\theta}^*; x, y) - \frac{1}{\sigma_{\mathrm{est}}} \sum_{\sigma=1}^{\sigma_{\mathrm{est}}} o^{\mathrm{loss}}_{\boldsymbol{\theta}^*,\sigma} \in \mathcal{O}\left(C_{\mathrm{loss}} \left(\sqrt{\frac{T \log(T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\frac{\log(1/\delta)}{\sigma_{\mathrm{est}}}} \right) \right). \tag{121}$$

Proof. We first insert a zero in terms of the empirical risk as follows:

$$\mathop{\mathbb{E}}_{x,y} \ell(\boldsymbol{\theta}^*; x, y) - \frac{1}{\sigma_{\mathrm{est}}} \sum_{\sigma=1}^{\sigma_{\mathrm{est}}} o^{\mathrm{loss}}_{\boldsymbol{\theta}^*,\sigma} = \left(\mathop{\mathbb{E}}_{x,y} \ell(\boldsymbol{\theta}^*; x, y) - \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{\theta}^*; x_i, y_i) \right) + \left(\frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{\theta}^*; x_i, y_i) - \frac{1}{\sigma_{\mathrm{est}}} \sum_{\sigma=1}^{\sigma_{\mathrm{est}}} o^{\mathrm{loss}}_{\boldsymbol{\theta}^*,\sigma} \right). \tag{122}$$

By Theorem 6, with probability at least $1 - \frac{\delta}{2}$ over the choice of the training data, the first term on the right-hand side (which is independent of the subsampling and of the obtained measurement outcomes) is bounded as $\in \mathcal{O}\left(C_{\mathrm{loss}} \left(\sqrt{\frac{T \log(T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} \right) \right)$. By Hoeffding's inequality, for any fixed S and $\boldsymbol{\theta}^*$, with probability at least $1 - \frac{\delta}{2}$ over the sampling of $i_1, \ldots, i_{\sigma_{\mathrm{est}}}$ and over the σ_{est} obtained measurement outcomes, the second term on the right-hand side is $\leqslant C_{\mathrm{loss}} \sqrt{\frac{2 \log(2/\delta)}{\sigma_{\mathrm{est}}}}$. Therefore, we also have

$$\mathbb{P}\left[\frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{\theta}^*; x_i, y_i) - \frac{1}{\sigma_{\mathrm{est}}} \sum_{\sigma=1}^{\sigma_{\mathrm{est}}} o^{\mathrm{loss}}_{\boldsymbol{\theta}^*,\sigma} > C_{\mathrm{loss}} \sqrt{\frac{2 \log(2/\delta)}{\sigma_{\mathrm{est}}}} \right] \tag{123}$$

$$= \mathbb{E}_{S,\boldsymbol{\theta}^*} \left[\mathbb{P}\left[\frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{\theta}^*; x_i, y_i) - \frac{1}{\sigma_{\mathrm{est}}} \sum_{\sigma=1}^{\sigma_{\mathrm{est}}} o^{\mathrm{loss}}_{\boldsymbol{\theta}^*,\sigma} > C_{\mathrm{loss}} \sqrt{\frac{2 \log(2/\delta)}{\sigma_{\mathrm{est}}}} \;\middle|\; S, \boldsymbol{\theta}^* \right] \right] \tag{124}$$

$$\leqslant \mathbb{E}_{S,\boldsymbol{\theta}^*} \left[\frac{\delta}{2} \right] \tag{125}$$

$$= \frac{\delta}{2}. \tag{126}$$

Now, the statement of the Corollary follows via a union bound. □

This shows that we do not need to perform a disproportionately large number of measurements to guarantee that the estimated training error is indeed a good proxy for the prediction error. It suffices to choose σ_{est} to be roughly $N/T \log(T)$, along with N being sufficiently larger than $T \log(T)$, to guarantee that the prediction error will not be much higher than the approximate (observed) training error.

7. *Mother theorem*

We can summarize all the previously discussed extensions of Theorem 6 in Theorem 4, which we restate here for convenience:

Theorem 4 (Mother Theorem)**.** Let $\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}}(\cdot)$ be a QMLM with a variable structure. Suppose that, for every $k \in \mathbb{N}$, there are at most $G_\tau \in \mathbb{N}$ allowed structures with exactly τ parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is taken from $\mathcal{M}_t$ and used M_t times, and an arbitrary number of non-trainable, global CPTP maps. Also, for each $t \in \mathbb{N}$, let $\mathcal{E}_t^0 \in \mathcal{CPTP}\left((\mathbb{C})^{\otimes 2}\right)$ be a fixed reference CPTP map. Let P be a probability distribution over input-output pairs. Suppose that, given training data $S = \{(x_i, y_i)\}_{i=1}^{N}$ of size N, our optimization of the QMLM over structures and parameters w.r.t. the loss function $\ell(\boldsymbol{\alpha}; x_i, y_i) = \mathrm{Tr}\left[O^{\mathrm{loss}}_{x_i,y_i} (\mathcal{E}_{\boldsymbol{\alpha}}^{\mathrm{QMLM}} \otimes \mathrm{id})(\rho(x_i)) \right]$ yields a (data-dependent) structure with $T = T(N)$ independently parameterized 2-qubit CPTP maps, in which the t^{th} of these maps is taken from $\mathcal{M}_t$ and used M_t times, as well as the parameter setting $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(S)$.

Then, with probability at least $1-\delta$ over the choice of i.i.d. training data S of size N according to P,

$$R(\boldsymbol{\alpha}^*) - \hat{R}_S(\boldsymbol{\alpha}^*)$$
$$\in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(M_t)}{N}} + \sum_{\substack{t=1\\ t\neq t_1,\ldots,t_K}}^{T} M_t\Delta_t + \sqrt{\frac{\log(G_T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right\}\right), \tag{127}$$

where $\Delta_1^T, \ldots, \Delta_T^T$ denote the (data-dependent) distance between the trainable maps in the output QMLM to the respective reference maps $\mathcal{E}_1^0, \ldots, \mathcal{E}_T^0$, $C_{\text{loss}} = \sup_{x,y} \|O_{x,y}^{\text{loss}}\|$ is the maximum (absolute) value attainable by the loss function, and the minimum is over all $K \in \{0, \ldots, T\}$ and choices of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, T\}$.

Moreover, if the loss is not evaluated exactly, but an unbiased estimator is built from σ_{est} subsampled training data points (as in Supplementary Note 3. 2. 6.), we only incur an additional error of $\mathcal{O}\left(\sqrt{\log(1/\delta)/\sigma_{\text{est}}}\right)$.

Proof. To prove this most general version of our results, we combine the previous results and proof strategies. First, fix $\tau \in \mathbb{N}$ and one of the G_τ admissible QMLM architectures with exactly τ trainable 2-qubit CPTP maps, in which the t^{th} of these maps is taken from $\mathcal{M}_t$ and used M_t times. With the same strategy as in the proof of Theorem 8, if we take $\mathcal{N}_t$ to be a (ε/KM_t)-covering net for $\mathcal{M}_t$ w.r.t. $\|\cdot\|_\diamond$, and consider the set of n-qubit CPTP maps obtained from the QMLM if exactly K of the τ independently trainable 2-qubit CPTP maps are taken from the respective $\mathcal{N}_t$, and the remaining $\tau - K$ maps are left at the corresponding reference map, this gives us an ε_K-covering net $\mathcal{N}_\varepsilon$ of the class of n-qubit CPTP maps that the QMLM architecture can implement, where

$$\varepsilon_K \coloneqq \varepsilon + \sum_{\substack{t=1\\ t\neq t_1,\ldots,t_K}}^{T} M_t\Delta_t. \tag{128}$$

This ε_K-covering net can be taken to have cardinality bounded as

$$\log(|\mathcal{N}_\varepsilon|) \leqslant K\log(\tau) + K \max_{1\leqslant t\leqslant \tau} c_t \log\left(1 + KM_t/\varepsilon\right). \tag{129}$$

If we use this metric entropy bound for the chaining argument presented in the proof of Theorem 7, we can show that, with probability $\geqslant 1 - \delta/2G_\tau\tau^2$ over the choice of i.i.d. training data S of size N, if $\boldsymbol{\theta}_\tau^*$ is a (continuous) parameter setting for the τ parameterized maps obtained through optimization upon data S, we have

$$R(\boldsymbol{\theta}_\tau^*) - \hat{R}_S(\boldsymbol{\theta}_\tau^*) \in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max\limits_{1\leqslant t\leqslant \tau} c_t \log(KM_t)}{N}} + \sqrt{\frac{K\log(\tau)}{N}} + \sum_{\substack{t=1\\ t\neq t_1,\ldots,t_K}}^{T} M_t\Delta_t + \sqrt{\frac{\log(2G_\tau\tau^2/\delta)}{N}}\right\}\right), \tag{130}$$

where the minimum is over $K \in \{0, \ldots, \tau\}$ and over choices of pairwise distinct $t_1, \ldots, t_K \in \{1, \ldots, \tau\}$.

We now first take a union bound over the G_τ admissible structures and then another union bound over $\tau \in \mathbb{N}$ to obtain: With probability $\geqslant 1-\delta$ over the choice of i.i.d. training data S of size N, if the optimization upon input of S outputs a QMLM architecture with $T = T(N)$ parameterized 2-qubit CPTP maps and the (discrete and continuous) parameter setting $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(S)$, then we have the generalization error bound

$$R(\boldsymbol{\alpha}^*) - \hat{R}_S(\boldsymbol{\alpha}^*) \tag{131}$$
$$\in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(KM_t)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sum_{\substack{t=1\\ t\neq t_1,\ldots,t_K}}^{T} M_t\Delta_t + \sqrt{\frac{\log(2G_T T^2/\delta)}{N}}\right\}\right) \tag{132}$$
$$\in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(K)}{N}} + \sqrt{\frac{K\log(T)}{N}} + \sqrt{\frac{K \max\limits_{1\leqslant t\leqslant T} c_t \log(M_t)}{N}} + \sum_{\substack{t=1\\ t\neq t_1,\ldots,t_K}}^{T} M_t\Delta_t + \sqrt{\frac{\log(G_T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right\}\right), \tag{133}$$

with the minimum as claimed.

To understand the added error in the case in which an unbiased estimate of the empirical risk is used, we now repeat the analysis given in Supplementary Note 3. 2. 6., but use the generalization bound just established instead of the one from Theorem 6, and obtain the claimed bound. □

Also for Theorem 4, we shortly explain how this leads to the result stated as Theorem 5 in the main text. First, Theorem 5 only considers the case $M_t = M$ for all t. Second, just like presented in Supplementary Note 3. 2. 5., we can bound the constants c_t by their worst-case upper bound of 1024, and then take a minimum not over all K and all choices of $t_1, \ldots, t_T$, but only over all K with the fixed choice $t_k = k$. With these two simplifications, the bound becomes

$$R(\boldsymbol{\alpha}^*) - \hat{R}_S(\boldsymbol{\alpha}^*) \in \mathcal{O}\left(C_{\text{loss}} \min\left\{\sqrt{\frac{K\log(MT)}{N}} + \sum_{\substack{t=1 \\ t\neq t_1,\ldots,t_K}}^{T} M\Delta_t + \sqrt{\frac{\log(G_T)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right\}\right).$$

Finally, once we plug in a constant confidence level δ and also consider C_{loss} as a constant dictated by the problem, we end up with the bound of Theorem 5 from the main text.

Remark 8. In Theorem 4, we have chosen a fixed reference map $\mathcal{E}_t^0 \in \mathcal{CPTP}\left((\mathbb{C})^{\otimes 2}\right)$ for every $t \in \mathbb{N}$. One could even choose different reference maps for each k and for each of the G_k allowed structures with exactly k parameterized maps.

In Supplementary Note 3. 2. 5., we have taken the initial point of the optimization procedure as reference point for evaluating distances. A similar interpretation is possible in Theorem 4, however, the reference points can be more abstract. In principle, the reference maps can be chosen freely, as long as the choice is independent of the training data sample w.r.t. which the empirical risk is evaluated.

Remark 9. We present our results for the case of a QMLM $\mathcal{E}_{\boldsymbol{\alpha}}^{\text{QMLM}}(\cdot)$ acting on a quantum input state $\rho(x)$. If x describes classical data, this presumes an "encoding-first" architecture, in which the classical-to-quantum data-encoding $x \mapsto \rho(x)$ is applied first, followed by a trainable quantum circuit. As observed in [22, 55, 56], the expressive power of a QMLM for processing classical data can significantly benefit from allowing for data re-uploading [57]. This is achieved by allowing for a more flexible form of QMLM, in which data-encoding and trainable gates can be interleaved. Our results, which focus on the trainable part of the QMLM circuit, directly extend to QMLMs with data re-uploading.

This can be seen as follows: In our proofs of the metric entropy bounds from Subsection 3. 1., we already allowed for an interleaving of the trainable gates with arbitrary fixed gates. The same reasoning applies if we replace the fixed gates by encoding gates depending on the classical input data x, as long as they are still independent of the trainable parameters.

Supplementary Note 4. Application to quantum phase recognition

As a second application of our prediction error bounds, we demonstrate their implications for quantum phase recognition (QPR) with quantum convolutional neural networks (QCNNs). Here, for each training example $(|\psi_i\rangle, y_i)$, the encoded input is simply $\rho(x_i) = |\psi_i\rangle\langle\psi_i|$, a pure n-qubit quantum state that belongs to one of four possible quantum phases of matter. The corresponding output label $y_i \in \{0,1\}^2$ tells us to which of the four phases $\rho(x_i)$ belongs. The goal of a quantum machine learning model for this scenario is to accurately predict, given a new input x, the corresponding label, and thus the phase, of the state ψ_i.

In our language, a QCNN acting on n-qubit states, as introduced in [25], is a QMLM $\mathcal{E}_{\boldsymbol{\theta}}^{\text{QCNN}}(\cdot)$ with a particular fixed structure, explained in more detail in Section II.C of the main text, consisting of $\log(n)$ independently parameterized 2-qubit maps, each of which is used at most n times. By measuring some of the qubits and then discarding them in pooling layers, the QCNN maps an n-qubit input to a 2-qubit output, on which it then performs a computational basis measurement. The phase prediction that the QCNN makes for an n-qubit input state is the one corresponding to the smallest of the four outcome probabilities in the computational basis measurement on the output state. This can be well approximated by running multiple gate-sharing copies of the QCNN in parallel and appropriately post-processing the single measurement outcomes. For simplicity of presentation, we showcase our bounds in the scenario of only one copy of the QCNN. However, this extends to multiple gate-sharing copies according to Corollary 2. Thus, we consider the loss function characterized by the loss observables

$$O_{x_i,y_i}^{\text{loss}} = O_{y_i}^{\text{loss}} = |y_i\rangle\langle y_i|, \tag{134}$$

which is independent of x_i. This means the loss function is given by

$$\ell(\boldsymbol{\theta}; \psi_i, y_i) \coloneqq \langle y_i|\mathcal{E}_{\boldsymbol{\theta}}^{\text{QCNN}}(|\psi_i\rangle\langle\psi_i|)|y_i\rangle. \tag{135}$$

That is, the QMLM achieves a small value of the loss function on the example $(|\psi_i\rangle, y_i)$ if the probability observing y_i when performing a computational basis measurement on the output state, upon input $|\psi_i\rangle$, is small. Correspondingly, the true risk is

$$R(\boldsymbol{\theta}) = \mathbb{E}_{(\psi_i,y_i)\sim P}[\langle y|\mathcal{E}^{\mathrm{QCNN}}_{\boldsymbol{\theta}}(|\,\psi\rangle\langle\psi\,|)|y\rangle] \tag{136}$$

and the empirical risk on training data $S = \{(|\psi_i\rangle, y_i)\}_{i=1}^N$ is

$$\hat{R}_S(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\langle y_i|\mathcal{E}^{\mathrm{QCNN}}_{\boldsymbol{\theta}}(|\,\psi_i\rangle\langle\psi_i\,|)|y_i\rangle. \tag{137}$$

With the scenario established, we can now apply the prediction error bound proved in Corollary 2. Here, it takes the form: Suppose that, given training data S of size N, our optimization of the parameters in the QCNN yields a parameter setting $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(S)$. Then, with probability $\geqslant 1-\delta$ over the choice of training data,

$$\mathbb{E}_{(\psi_i,y_i)\sim P}[\langle y|\mathcal{E}^{\mathrm{QCNN}}_{\boldsymbol{\theta}}(|\,\psi\rangle\langle\psi\,|)|y\rangle] \leqslant \frac{1}{N}\sum_{i=1}^{N}\langle y_i|\mathcal{E}^{\mathrm{QCNN}}_{\boldsymbol{\theta}}(|\,\psi_i\rangle\langle\psi_i\,|)|y_i\rangle + \mathcal{O}\left(\sqrt{\frac{\log(n)^2+\log(1/\delta)}{N}}\right). \tag{138}$$

Therefore, a small training error guarantees a small prediction error already for training data size $N \in \mathcal{O}(\mathrm{poly}(\log(n)))$. In other words, when using a QCNN for QPR, a good generalization error is already guaranteed for training data of size poly-logarithmic in n, the number of qubits. Thereby, our results provide a rigorous explanation for the good generalization behavior of QCNNs even for small training data size that was observed numerically in [25].

Supplementary Note 5. Application to unitary compiling

The second application of our generalization guarantees to be presented here is that of learning unitaries in the sense of (quantum-assisted) unitary compiling [38]. Unitary compiling is the task of finding a circuit representation of a target unitary, given black-box access to that unitary.

From a learning perspective, this motivates the following problem: For each training example (x_i, y_i), the input is a pure n-qubit state $\rho(x_i) = |\psi_i\rangle\langle\psi_i|$, and the corresponding label is the pure n-qubit state $|\phi_i\rangle\langle\phi_i| = U|\psi_i\rangle\langle\psi_i|U^\dagger$ obtained by unitarily evolving the input state according to the (unknown) target unitary U. We consider the loss function given induced by the trace distance via

$$\ell(\boldsymbol{\alpha}; |\psi\rangle, |\phi\rangle) \coloneqq \|\,|\phi\rangle\langle\phi| - \mathcal{U}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}}(|\,\psi\rangle\langle\psi\,|)\|_1^2. \tag{139}$$

where $\mathcal{U}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}}(\cdot) = U_{\boldsymbol{\alpha}}(\cdot)U^\dagger_{\boldsymbol{\alpha}}$ is a (unitary) quantum machine learning model, and we take $|\phi\rangle = U|\psi\rangle$.

As we are considering a trace distance between pure states, we can rewrite the loss function in terms of the fidelity (i.e., the overlap) as

$$\ell(\boldsymbol{\alpha}; |\psi\rangle, |\phi\rangle) = 1 - |\langle\phi|U_{\boldsymbol{\alpha}}\psi\rangle|^2 = 1 - \mathrm{Tr}\left[|\,\phi\rangle\langle\phi\,| \cdot \mathcal{U}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}}(|\,\psi\rangle\langle\psi\,|)\right]. \tag{140}$$

Hence, we see that this loss function is encompassed by our scenario, because we can write

$$\ell(\boldsymbol{\alpha}; |\psi\rangle, |\phi\rangle) = \mathrm{Tr}\left[O^{\mathrm{loss}}_{\psi,\phi} \cdot \mathcal{U}^{\mathrm{QMLM}}_{\boldsymbol{\alpha}}(|\,\psi\rangle\langle\psi\,|)\right], \tag{141}$$

with loss observables $O^{\mathrm{loss}}_{\psi,\phi} = \mathbb{1} - |\,\phi\rangle\langle\phi\,|$ (depending only on the quantum output, but not on the input).

With (the above rewriting of) this loss function, the expected loss, when the expectation is w.r.t. drawing the input states independently at random from the Haar measure, becomes connected to the Hilbert-Schmidt inner product between the target unitary and the unitary implemented by the circuit. This, in turn, can be given an operational interpretation, as detailed in [38].

We solve this learning problem using a QMLM with a variable structure. (See Sections II.C and IV. of the main text for more details on how this is implemented.) In this scenario, Corollary 3 implies that, if we optimize over both (discrete) structures and (continuous) parameters and obtain an output structure $\boldsymbol{k}^*$ with T parameterized gates with a parameter setting $\boldsymbol{\alpha}^* = (\boldsymbol{\theta}^*, \boldsymbol{k}^*)$, then, with probability $\geqslant 1-\delta$ over the choice of training data of size N, which is drawn i.i.d. from some distribution P over pure n-qubit states, we are guaranteed that

$$\mathbb{E}_{|\psi\rangle\sim P}\left[\|U|\,\psi\rangle\langle\psi\,|U^\dagger - U_{\boldsymbol{\alpha}^*}|\,\psi\rangle\langle\psi\,|U^\dagger_{\boldsymbol{\alpha}^*}\|_1^2\right] \leqslant \frac{1}{N}\sum_{i=1}^{N}\|U|\,\psi_i\rangle\langle\psi_i\,|U^\dagger - U_{\boldsymbol{\alpha}^*}|\,\psi_i\rangle\langle\psi_i\,|U^\dagger_{\boldsymbol{\alpha}^*}\|_1^2 + \tilde{\mathcal{O}}\left(\sqrt{\frac{T}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right), \tag{142}$$

assuming that the number of allowed structures with T gates scales at most exponentially in T. Here, the $\tilde{\mathcal{O}}$ hides terms logarithmic in T.

Consequently, we know that, with high probability, the trace distance between the state obtained by applying the learned unitary on a new unseen input state (drawn at random from the data-generating distribution) and the true output state will be small if we can achieve a small average trace distance over the N randomly sampled states, where N scales roughly linearly in T. For many unitary gates of interest, namely those that can be efficiently implemented, we thus expect T, and thereby also N, to scale polynomially in n, the number of qubits. This is a substantial improvement over the training data sizes used in previous approaches to unitary compiling, which were often taken to be exponential in n such as to uniquely determine the unknown target unitary [36, 37, 42]. This improvement comes at the cost of not compiling the target unitary exactly, but only with a certain (small) accuracy and success probability. Nevertheless, for many applications, paying this cost is worthwhile, given the significant savings in training data size guaranteed by our results.

As a concrete example, the QFT discussed in Section II.C of the main text can be exactly implemented with $T \in \mathcal{O}(n^2)$ gates. In this case, our theory implies that $N \in \mathcal{O}(n^2)$ training data points are, with high probability, sufficient for good generalization. As discussed in [58], approximate implementations of the QFT are possible with a lower number of gates, namely with $T \in \mathcal{O}(n \log(n))$. Potentially, one can combine this insight with our result to obtain a similar improvement in the upper bound on the sufficient training data size.

[1] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," Th. Prob. App. **16**, 264–280 (1971).
[2] David Pollard, *Convergence of stochastic processes* (Springer, 1984).
[3] Evarist Giné and Joel Zinn, "Some limit theorems for empirical processes," The Annals of Probability , 929–989 (1984).
[4] Peter L Bartlett and Shahar Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," Journal of Machine Learning Research **3**, 463–482 (2002).
[5] Richard M. Dudley, *Uniform Central Limit Theorems* (Cambridge University Press, 1999).
[6] Olivier Bousquet and André Elisseeff, "Stability and generalization," The Journal of Machine Learning Research **2**, 499–526 (2002).
[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference* (Springer, 2006) pp. 265–284.
[8] Nick Littlestone and Manfred Warmuth, "Relating data compression and learnability," Technical report, University of California Santa Cruz (1986).
[9] David A McAllester, "Some pac-bayesian theorems," Machine Learning **37**, 355–363 (1999).
[10] Matthias C. Caro and Ishaun Datta, "Pseudo-dimension of quantum circuits," Quantum Machine Intelligence **2**, 14 (2020).
[11] Claudiu Marius Popescu, "Learning bounds for quantum circuits in the agnostic setting," Quantum Information Processing **20**, 1–24 (2021).
[12] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner, "The power of quantum neural networks," Nature Computational Science **1**, 403–409 (2021).
[13] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang, "Statistical complexity of quantum circuits," Physical Review A **105**, 062431 (2022).
[14] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang, "Effects of quantum resources on the statistical complexity of quantum circuits," arXiv preprint arXiv:2102.03282 (2021).
[15] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang, "Rademacher complexity of noisy quantum circuits," arXiv preprint arXiv:2103.03139 (2021).
[16] Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao, "Efficient measure for the expressivity of variational quantum algorithms," Physical Review Letters **128**, 080506 (2022).
[17] Leonardo Banchi, Jason Pereira, and Stefano Pirandola, "Generalization in quantum machine learning: A quantum information standpoint," PRX Quantum **2**, 040321 (2021).
[18] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean, "Power of data in quantum machine learning," Nature Communications **12**, 1–9 (2021).
[19] Xinbiao Wang, Yuxuan Du, Yong Luo, and Dacheng Tao, "Towards understanding the power of quantum kernels in the nisq era," Quantum **5**, 531 (2021).
[20] Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf, "The inductive bias of quantum kernels," Advances in Neural Information Processing Systems **34**, 12661–12673 (2021).
[21] Casper Gyurik, Dyon van Vreumingen, and Vedran Dunjko, "Structural risk minimization for quantum linear classifiers," arXiv preprint arXiv:2105.05566 (2021).
[22] Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke, "Encoding-dependent generalization bounds for parametrized quantum circuits," Quantum **5**, 582 (2021).
[23] Chih-Chieh Chen, Masaya Watabe, Kodai Shiba, Masaru Sogabe, Katsuyoshi Sakamoto, and Tomah Sogabe, "On the expressibility and overfitting of quantum circuit learning," ACM Transactions on Quantum Computing **2**, 1–24 (2021).
[24] Haoyuan Cai, Qi Ye, and Dong-Ling Deng, "Sample complexity of learning parametric quantum circuits," Quantum Science and Technology **7**, 025014 (2022).
[25] Iris Cong, Soonwon Choi, and Mikhail D Lukin, "Quantum convolutional neural networks," Nature Physics **15**, 1273–1278 (2019).
[26] Frank Schindler, Nicolas Regnault, and Titus Neupert, "Probing many-body localization with neural networks," Phys. Rev. B **95**, 245134 (2017).
[27] Lei Wang, "Discovering phase transitions with unsupervised learning," Phys. Rev. B **94**, 195105 (2016).
[28] Evert P. L. van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber, "Learning phase transitions by confusion," Nature Physics **13**, 435 (2017).
[29] Juan Carrasquilla and Roger G. Melko, "Machine learning phases of matter," Nature Physics **13**, 431 (2017).
[30] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven, "Barren plateaus in quantum neural network training landscapes," Nature Communications **9**, 1–6 (2018).
[31] M Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J Coles, "Cost function dependent barren plateaus in shallow parametrized quantum circuits," Nature Communications **12**, 1–12 (2021).
[32] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill, "Provably efficient machine learning for quantum many-body problems," arXiv preprint arXiv:2106.12627 (2021).
[33] Davide Venturelli, Minh Do, Eleanor Rieffel, and Jeremy Frank, "Compiling quantum circuits to realistic hardware architectures using temporal planners," Quantum Science and Technology **3**, 025004 (2018).
[34] Kyle Booth, Minh Do, J Beck, Eleanor Rieffel, Davide Venturelli, and Jeremy Frank, "Comparing and integrating constraint programming and temporal planning for quantum circuit compilation," in *Proceedings of the International Conference on*

Automated Planning and Scheduling, Vol. 28 (2018).
[35] Keri A McKiernan, Erik Davis, M Sohaib Alam, and Chad Rigetti, "Automated quantum programming via reinforcement learning for combinatorial optimization," arXiv preprint arXiv:1908.08054 (2019).
[36] Lukasz Cincio, Yiğit Subaşı, Andrew T Sornborger, and Patrick J Coles, "Learning the quantum algorithm for state overlap," New Journal of Physics **20**, 113022 (2018).
[37] Lukasz Cincio, Kenneth Rudinger, Mohan Sarovar, and Patrick J. Coles, "Machine learning of noise-resilient quantum circuits," PRX Quantum **2**, 010324 (2021).
[38] Sumeet Khatri, Ryan LaRose, Alexander Poremba, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles, "Quantum-assisted quantum compiling," Quantum **3**, 140 (2019).
[39] Kunal Sharma, Sumeet Khatri, M. Cerezo, and Patrick J Coles, "Noise resilience of variational quantum compiling," New Journal of Physics **22**, 043006 (2020).
[40] Kentaro Heya, Yasunari Suzuki, Yasunobu Nakamura, and Keisuke Fujii, "Variational quantum gate optimization," arXiv preprint arXiv:1810.12745 (2018).
[41] Tyson Jones and Simon C Benjamin, "Robust quantum compilation and circuit optimisation via energy minimisation," Quantum **6**, 628 (2022).
[42] E. Younis and L. Cincio, "Quantum Fast Circuit Optimizer (qFactor)," .
[43] Kunal Sharma, Marco Cerezo, Zoë Holmes, Lukasz Cincio, Andrew Sornborger, and Patrick J Coles, "Reformulation of the no-free-lunch theorem for entangled datasets," Physical Review Letters **128**, 070501 (2022).
[44] M Bilkis, M Cerezo, Guillaume Verdon, Patrick J Coles, and Lukasz Cincio, "A semi-agnostic ansatz with variable structure for quantum machine learning," arXiv preprint arXiv:2103.06712 (2021).
[45] M. Bilkis, "An implementation of VAns: A semi-agnostic ansatz with variable structure for quantum machine learning," .
[46] Wassily Hoeffding, "Probability inequalities for sums of bounded random variables," Journal of the American Statistical Association **58**, 13–30 (1963).
[47] Colin McDiarmid, "On the method of bounded differences," in *Surveys in combinatorics, 1989 (Norwich, 1989)*, London Math. Soc. Lecture Note Ser., Vol. 141 (Cambridge Univ. Press, Cambridge, 1989) pp. 148–188.
[48] Pascal Massart, "Some applications of concentration inequalities to statistics," Annales de la Faculté des sciences de Toulouse : Mathématiques **Ser. 6, 9**, 245–303 (2000).
[49] John Watrous, *The Theory of Quantum Information* (Cambridge University Press, 2018).
[50] Chi-Fang Chen, Hsin-Yuan Huang, Richard Kueng, and Joel A Tropp, "Concentration for random product formulas," PRX Quantum **2**, 040305 (2021).
[51] Christopher A Fuchs and Jeroen Van De Graaf, "Cryptographic distinguishability measures for quantum-mechanical states," IEEE Transactions on Information Theory **45**, 1216–1227 (1999).
[52] Hsin-Yuan Huang, Richard Kueng, and John Preskill, "Information-theoretic bounds on quantum advantage in machine learning," Phys. Rev. Lett. **126**, 190505 (2021).
[53] Roman Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge University Press, 2018).
[54] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of Machine Learning* (MIT Press, 2018).
[55] Francisco Javier Gil Vidal and Dirk Oliver Theis, "Input redundancy for parameterized quantum circuits," Frontiers in Physics **8**, 297 (2020).
[56] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models," Physical Review A **103**, 032430 (2021).
[57] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre, "Data re-uploading for a universal quantum classifier," Quantum **4**, 226 (2020).
[58] Yunseong Nam, Yuan Su, and Dmitri Maslov, "Approximate quantum fourier transform with o (n log (n)) t gates," NPJ Quantum Information **6**, 1–6 (2020).

Appendix C

Further preprints and articles as principal author under review

C.1 Undecidability of Learnability

Undecidability of Learnability

Matthias C. Caro

In many classical learning scenarios, we know characterizations of learnability in terms of combinatorial properties of the hypothesis class. These results already give important insights into learnability because they allow us to translate abstract questions of learnability to concrete combinatorial properties. This, however, now raises the question: How challenging is it to determine these combinatorial properties? In this work, we answer this question from the perspective of formal logic and computability.

We begin the article by motivating our work, summarizing our main results, and putting them into the context of prior work. This is followed by our results on the undecidability of finiteness of the VC dimension, and thus of probably approximately correct learnability in binary classification tasks, in Section 2. Namely, in Subsection 2.2, we show how to construct, given a recursively enumerable and consistent formal system F, a function class $\mathcal{G}_F$ that has finite VC dimension, but for which F cannot prove this finiteness. In this work, we call such a true but unprovable statement Gödel undecidable. In a similar spirit, in Subsection 2.3 we show Turing undecidability of finiteness of the VC dimension. That is, we show that there is no algorithm that, given the code of a computable hypothesis class, decides whether that class has finite VC dimension. We achieve this by constructing in a computable way, given the code of a Turing machine M, a computable function class $\mathcal{H}_M$ that has finite VC dimension if and only if M halts on the empty input.

In Section 3, we establish analogous undecidability results for scenarios modelling teacher-learner interactions. Here, the relevant combinatorial parameter is the so-called teaching dimension. In Subsection 3.3, we show Gödel undecidability for finiteness of the teaching dimension of $\mathcal{G}_F$. We follow this in Subsection 3.4 by proving Turing undecidability for finiteness of the teaching dimension, again using the computable mapping $M \mapsto \mathcal{H}_M$. In addition to these two results, we also show in Subsection 3.2 that a related problem, namely that of deciding whether a function admits a finite teaching set, is undecidable in the same two senses.

Section 4 extends our discussion to also include online learning. More precisely, we show that both uniform and universal online learnability are Gödel and Turing undecidable. For uniform online learning, we achieve this by proving undecidability of finiteness of the Littlestone dimension. In the case of universal online learning, we argue via the (non-)existence of infinite Littlestone trees.

The idea for this project was motivated by discussions between my doctoral advisor, Michael M. Wolf, and myself. In particular, he suggested the construction used in Subsection 3.2 of the article. Also, since making the first version of the work available on the arXiv, I have included some comments suggested by other researchers, as detailed in the acknowledgements of the paper. I am solely responsible for the scientific content of this article, with the two restrictions just mentioned. As the single author of this article, I am solely responsible for writing this article.

Permission to include:

Matthias C. Caro.
Undecidability of Learnability.
arXiv preprint arXiv:2106.01382.

arXiv.org - Non-exclusive license to distribute

The URI http://arxiv.org/licenses/nonexclusive-distrib/1.0/ is used to record the fact that the submitter granted the following license to arXiv.org on submission of an article:

- I grant arXiv.org a perpetual, non-exclusive license to distribute this article.
- I certify that I have the right to grant this license.
- I understand that submissions cannot be completely removed once accepted.
- I understand that arXiv.org reserves the right to reclassify or reject any submission.

Revision history

2004-01-16 - License above introduced as part of arXiv submission process
2007-06-21 - This HTML page created

Contact

Undecidability of Learnability

Matthias C. Caro *
Technical University of Munich, Department of Mathematics, Garching, Germany
Munich Center for Quantum Science and Technology (MCQST), Munich, Germany
caro@ma.tum.de

Abstract

Machine learning researchers and practitioners steadily enlarge the multitude of successful learning models. They achieve this through in-depth theoretical analyses and experiential heuristics. However, there is no known general-purpose procedure for rigorously evaluating whether newly proposed models indeed successfully learn from data.

We show that such a procedure cannot exist. For PAC binary classification, uniform and universal online learning, and exact learning through teacher-learner interactions, learnability is in general undecidable, both in the sense of independence of the axioms in a formal system and in the sense of uncomputability. Our proofs proceed via computable constructions of function classes that encode the consistency problem for formal systems and the halting problem for Turing machines into complexity measures that characterize learnability. Our work shows that undecidability appears in the theoretical foundations of machine learning: There is no one-size-fits-all algorithm for deciding whether a machine learning model can be successful. We cannot in general automatize the process of assessing new learning models.

* *ORCiD:* 0000-0001-9009-2372

1 Introduction

One of the foundational questions in machine learning theory is "When is learning possible?" This is the question for necessary and sufficient conditions for learnability. Such conditions have been identified for different learning models. They can take the form of requiring a certain, often combinatorial, complexity measure to be finite. Well known examples of such complexity measures include the VC-dimension for binary classification in the PAC model, the Littlestone dimension for online learning, or different notions of teaching dimensions for teacher-learner interactions.

We consider a question that is slightly different from, but arguably just as important as the one above. Namely, we ask "Can we decide whether learning is possible?" At first glance, the ability to answer the first question might also seem to allow to resolve this second one. If, e.g., you know a complexity measure whose finiteness is equivalent to learnability, that gives you a criterion to decide learnability. However, whether this is indeed a satisfactory criterion strongly depends on the exact meaning of "decide" in the second question.

We consider two such meanings and thereby obtain two variants of the second question. The first is natural from a mathematician's perspective, namely "If a class is learnable, can we prove that this is the case?" The second is intimately familiar to computer scientists, namely "Does there exist an algorithm that decides learnability?" After specifying in either of these two ways what it means to "decide whether learning is possible," we see that the answer to the second question is not trivially positiveEven given the definition of a complexity parameter that is finite if and only if learning is possible, answering the second question still requires a proof of finiteness of that complexity measure or an algorithm that decides whether the complexity measure is finite or not.

In fact, we show that the answer to the question "Can we decide whether learning is possible?" is, in general, negative for both of the variants introduced above and for different learning scenarios. In particular, we demonstrate this for learning models in which criteria for learnability in terms of complexity measures are known. More concretely, we consider binary classification, uniform and universal online learning, and the task of exactly identifying a function through teacher-learner interactions. We show in all these scenarios: On the one hand, there is a function class that is learnable but whose learnability cannot be proved. On the other hand, there is no general-purpose algorithm that, upon input of a class, decides whether it is learnable.

1.1 Overview Over the Results

Our undecidability results come in two flavours, one about provability in a formal system, the other about computability via Turing machines. We summarize our line of reasoning in Figure 1 and explain it in more detail in the following paragraphs.

We first study binary classification in *Probably Approximately Correct (PAC)* learning. The relevant complexity measure for this learning scenario is the VC-dimension due to [VC71]. On the one hand, given a recursively enumerable formal system F, we define a class $\mathcal{G}_F \subseteq \{0,1\}^{\mathbb{N}}$

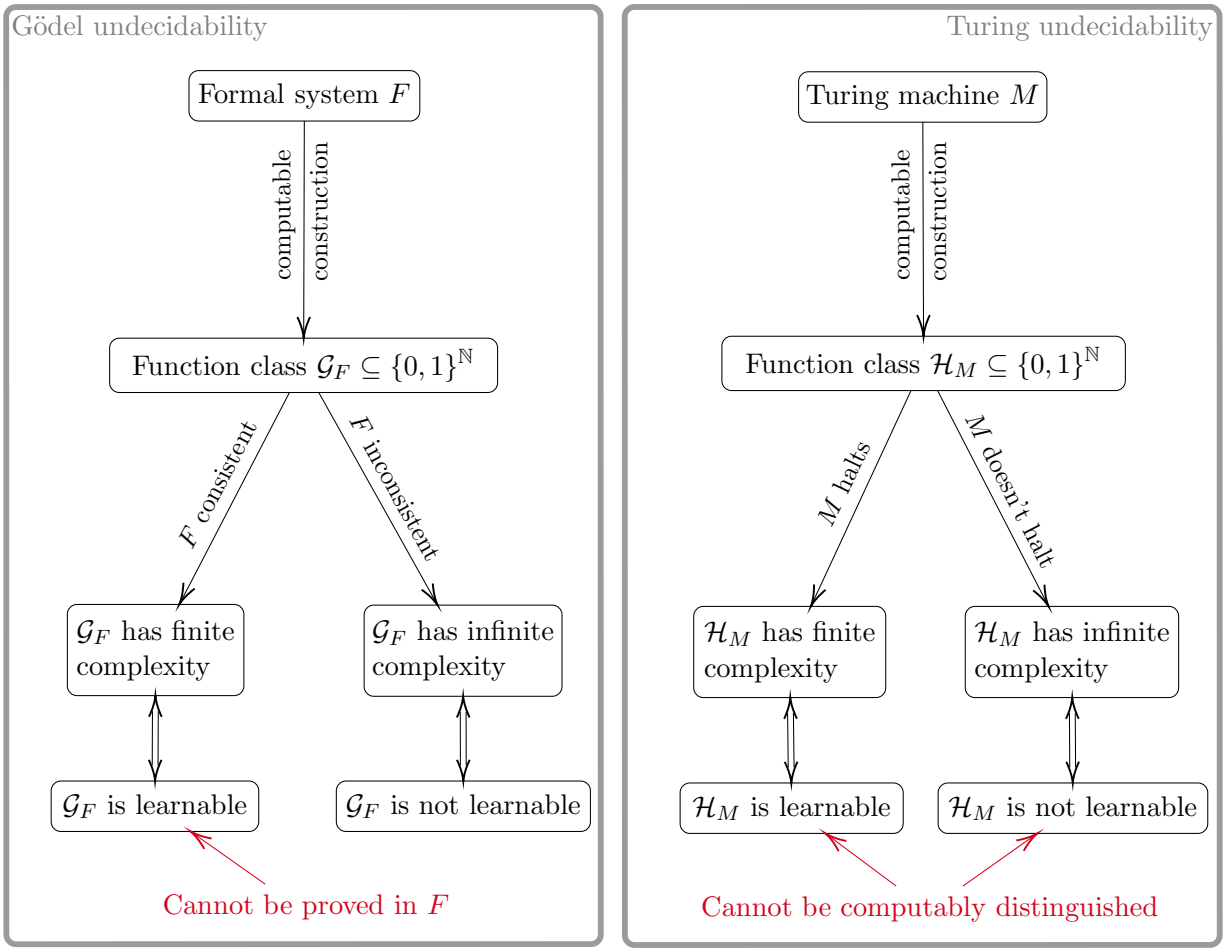


Figure 1: A depiction of our line of reasoning. "Complexity" is to be understood in terms of VC-dimension, teaching dimension, Littlestone dimension, or Littlestone trees, depending on the learning model. To conclude undecidability, we use Gödel's second incompleteness theorem and the uncomputability of the halting problem, respectively.

(Definition 2.4) that is PAC learnable if and only if F is consistent (Corollary 2.9). If Gödel's second incompleteness theorem applies to F, we conclude that the function class $\mathcal{G}_F$ is PAC learnable, but its PAC learnability cannot be proved in F (Corollary 2.11). On the other hand, given a Turing machine M, we define a class $\mathcal{H}_M \subseteq \{0,1\}^{\mathbb{N}}$ (Definition 2.15) that is PAC learnable if and only if M halts on the empty input (Lemma 2.16). By reduction to the halting problem, there is no general-purpose algorithm that decides whether a computable binary-valued function class is PAC learnable (Corollary 2.20).

Our constructions start from the recursively enumerable set $c_c(\{0,1\})$ of functions with compact support in $\mathbb{N}$. Depending on the underlying object, i.e., the formal system or the Turing machine, we then further restrict the function class. We implement these restrictions based on consistency of finitely many provable theorems and halting after finitely many steps. Thereby, we ensure that they are computable from the underlying object (see Corollaries 2.14 and 2.19). For the Gödel scenario,

this translates the assumption of the existence of a recursive enumeration of the provable theorems to a property of the function class. For the Turing scenario, this computability is necessary for a reduction to the halting problem.

We also consider $\mathcal{G}_F$ and $\mathcal{H}_M$ in the following scenario: A teacher can provide examples of a target function to help a learner identify that function. Here, the basic complexity measure is the teaching dimension [GK95], which is finite if and only if the teaching problem can be solved with finitely many examples. We show that $\mathcal{G}_F$ has finite teaching dimension if and only if F is consistent (Proposition 3.5). In this case, the teaching problem can be solved but this cannot proved in F (Corollary 3.6), assuming again that Gödel's second incompleteness theorem applies. Similarly, we show that $\mathcal{H}_M$ has finite teaching dimension if and only if the underlying Turing machine M halts on the empty input (Proposition 3.8). So, there is no algorithm for deciding whether a function class can be taught/learned (Corollary 3.9).

We demonstrate the undecidability of one more decision problem motivated by teacher-learner interactions. Namely, in general, one cannot decide whether a given function in a known class can be taught/learned from finitely many examples. Again, this is true both in the sense of independence of the axioms of a formal system (Corollary 3.3) and in the sense of uncomputability (Remark 3.4).

Finally, our constructions also yield undecidability results for uniform and universal online learning. For online learning with uniform mistake bounds, the Littlestone dimension [Lit88] is the corresponding complexity parameter. For universal online learning, the relevant complexity condition is whether there exist infinite Littlestone trees [Bou+20]. After showing (in Propositions 4.5, 4.7, 4.9, and 4.11) that whether these complexity conditions are satisfied by $\mathcal{G}_F$ and $\mathcal{H}_M$ is again determined by whether F is consistent and whether M halts on the empty input, respectively, we conclude: Both uniform and universal online learnability are, in general, both Gödel and Turing undecidable (Corollaries 4.6, 4.8, 4.10, and 4.12).

Compared to prior work on undecidability in learning theory, which we review in Subsection 1.2, our approach is at the same time more direct and is the first that simultaneously proves undecidability results for multiple established learning models both in the sense of formal independence and in the sense of uncomputability. Our main technical contribution consists in constructing and studying the function classes $\mathcal{G}_F$ and $\mathcal{H}_M$, which we base on a careful elaboration of the computational model. Conceptually, we show that many of the established learnability criteria in terms of complexity measures are undecidable, thus demonstrating a limitation of the approach towards learnability and model selection via such complexity measures.

1.2 Related Work

[Lat96] made an early investigation into the relationship between computability and learnability. The main question in [Lat96] is whether and under which notions of "learnability" one can consider an uncomputable problem to be learnable. More precisely, [Lat96] considered the task of learning the halting problem relative to an oracle.

Both [Sch99] and [Zha18] studied the computability of finiteness of the VC-dimension. In particular, Theorem 1 in [Zha18] and Theorem 4.1 in [Sch99] state: Deciding finiteness of the VC-dimension of a computable concept class is Σ_2-complete. This implies our Corollary 2.20, the Turing undecidability of finiteness of the VC-dimension. The proofs of [Sch99] and [Zha18] used that deciding finiteness of the domain of a computable function is Σ_2-complete (see, e.g., Theorem IV.3.2 in [Soa78]). [Zha18] additionally invoked a result by [Las92]: A function class uniformly definable via a first-order formula has finite VC-dimension if and only if the defining formula is an *NIP* formula. While one of our results is already implied by [Sch99] and [Zha18], we consider our work to be a significant extension in two directions: On the one hand, we consider both Turing and Gödel undecidability. On the other hand, our proof strategies are at the same time more direct, using no results from logic beyond Gödel's incompleteness theorems and the Turing undecidability of the halting problem, and flexibly applicable to other complexity measures and learning scenarios.

[Ben+19] proposed the "estimating-the-maximum" (EMX) problem and proved that learnability in this model is independent of the ZFC axioms. While this already indicates that learning can be undecidable, our results add new insight in at least two ways. First, our results are for already established learning models. In particular, whereas [Ben+19] showed that, assuming consistency of ZFC, there is no dimension-like quantity of finite character that characterizes EMX learnability, our results include scenarios in which such dimensions for learning exist. Second, whereas [Ben+19] used the continuum, the continuum hypothesis, and the axiom of choice, we only use natural numbers and computable objects. This allows us to prove uncomputability results, which cannot be derived from the results of [Ben+19]. Some implications and limitations of the approach of [Ben+19] have been discussed, e.g., in [Har19; Tay19; Gan20].

[Aga+20] initiated a study of computable learners, which then truly deserve to be called "learning algorithms." In particular, [Aga+20] showed that not every PAC learnable class admits a computable learner and also identified conditions under which PAC learnability implies copmutable PAC learnability. Thereby, [Aga+20] extended considerations from [Sol08], which studied the task of non-uniform learning over all computable functions by a computable learner. As the underlying questions of [Aga+20] and our work differ, the results are not comparable. However, as we show the function classes $\mathcal{G}_F$ and $\mathcal{H}_M$ to be computable, the results of [Aga+20] imply that our undecidability results hold not only for PAC learning, but also for computable PAC learning.

[SFM21] took yet another perspective on undecidabity in learning theory. Namely, [SFM21] considered the problem of deciding, given an algorithm $\mathcal{A}$ and a dataset d, whether $\mathcal{A}$ is a learning algorithm and the output model of $\mathcal{A}$ underfits d. Here, [SFM21] used an information-theoretic notion of underfitting. The main result of [SFM21]: This decision problem can be reduced to the halting problem and is thus Turing undecidable.

[Han+21] recently identified a further potential source of undecidability in learning theory. They studied the existence of universally Bayes consistent learners for countable multiclass classification, i.e., of learners whose classification error almost surely converges to the optimal Bayes risk (over all

Borel measurable classifiers) as the sample size goes to infinity. Theorem 4.1 in [Han+21] states: A universally Bayes consistent classifier can only exist if the metric space from which the instances are drawn is essentially separable. The existence of metric spaces that are not essentially separable, however, is believed to be independent of the ZFC axioms. Hence, whether all metric instance spaces admit a universally Bayes consistent classifier might turn out to be a learning-theoretic question independent of ZFC.

Combinatorial complexity measures for learning have also been studied from in computational complexity theory. [PY96], motivated by [LMR91], determined the complexity of computing the VC-dimension of a finite concept class over a finite domain. While they argue that this problem is probably not NP-complete, [PY96] proved its completeness for the complexity class LOGNP, a logarithmically-restricted version of NP. [Shi93] obtained a similar completeness result. [FL98] then, by reduction to computing the VC-dimension, established the LOGNP-hardness of computing the Littlestone dimension. [Sch99] later showed that a variant of the above problem, namely that of computing the VC-dimension of a class described by a polynomial-sized circuit, is Σ_3^p-complete. Computing the Littlestone dimension from a circuit description is PSPACE-complete [Sch00]. Extending a result by [Sch99], [MU02] determined the complexity of a promise version of approximating the VC-dimension of a class associated to a polynomial-size circuit. More recently, [MR17] has proved nearly tight quasi-polynomial time lower bounds for approximating the VC-dimension and the Littlestone dimension, assuming the randomized Exponential Time Hypothesis.

1.3 Structure of the Paper

Section 2 contains our main constructions and results leading to undecidability of finiteness of the VC-dimension. In Section 3, we demonstrate that our constructions also yield undecidability results for teaching problems. In Section 4, we exhibit analogous results for both uniform and universal online learning. We conclude with an outlook and open questions in Section 5. Full proofs appear either directly in the text or in Appendix A. Appendices B and C contain standard definitions and results related to formal systems and computability that are used in the main text.

2 Undecidability of Finiteness of the VC-Dimension

2.1 Preliminaries: PAC Binary Classification and the VC-Dimension

We start by recalling one of the most influential learning models for binary classification:

Definition 2.1 (Probably approximately correct binary classification [Val84])**.** *Let $\mathcal{X}$ be some space, write $\mathcal{Z} = \mathcal{X} \times \{0,1\}$. Let $\mathcal{G} \subset \{0,1\}^{\mathcal{X}}$, and let D be a probability distribution on $\mathcal{Z}$. A map $\mathcal{A} : \bigcup_{m=1}^{\infty} \mathcal{Z}^m \to \{0,1\}^{\mathcal{X}}$, $S \mapsto h_S$, is a* probably approximately correct (PAC) *learner for $\mathcal{G}$ if there exists a function $m : (0,1)^2 \to \mathbb{N}_{\geq 1}$ such that, given $\varepsilon, \delta \in (0,1)$, if $m \geq m(\varepsilon, \delta)$, then, with*

probability $\geq 1-\delta$ *with respect to repeated sampling of* $S \sim D^m$*, it holds that*

$$\mathbb{P}_{(x,y)\sim D}[h_S(x) \neq y] \leq \varepsilon + \inf_{g\in\mathcal{G}} \mathbb{P}_{(x,y)\sim D}[g(x) \neq y].$$

The PAC learners of interest are *polynomial PAC learners*, for which the sample size $m(\varepsilon, \delta)$ can be chosen to depend polynomially on $1/\varepsilon$ and $\log(1/\delta)$. Here, the "polynomial" refers to the sample size only, not to the runtime. If $\mathcal{G}$ admits a polynomial PAC learner, we call $\mathcal{G}$ *PAC learnable*.

For the scenario of binary classification, whether there exists a polynomial PAC learner can be understood in terms of a combinatorial quantity of the function class under consideration.

Definition 2.2 (VC-dimension [VC71])**.** *Let* $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$*. The Vapnik-Chervonenkis dimension, abbreviated as VC-dimension, of* $\mathcal{G}$ *is defined to be*

$$\mathrm{VCdim}(\mathcal{G}) := \sup\{n \in \mathbb{N}_0 \mid \exists S \subseteq \mathcal{X} : |S| = n \ \wedge \ |\mathcal{G}|_S| = 2^n\}.$$

If $S \subseteq \mathcal{X}$ *is a set such that* $|\mathcal{G}|_S| = 2^{|S|}$*, we say that* S *is shattered by* $\mathcal{G}$*.*

Under suitable measurability assumptions on the function class $\mathcal{G}$, we have the following

Theorem 2.3 (Fundamental theorem of binary classification (see, e.g., [SB19]))**.** *Let* $\mathcal{G} \subset \{0,1\}^{\mathcal{X}}$*.* $\mathcal{G}$ *is PAC learnable if and only if* $\mathrm{VCdim}(\mathcal{G}) < \infty$*.*

Among the assumptions on $\mathcal{G}$ that guarantee the equivalence of Theorem 2.3 are that $\mathcal{G}$ be image admissible Suslin, universally separable, well behaved, or countable [Dud78; Pol84; Blu+89; Pes11]. The function classes considered in this paper are all countable, so Theorem 2.3 applies. Therefore, when studying PAC learnability, we focus on studying finiteness of the VC-dimension. Interestingly, [AW18] found the VC-dimension to characterize quantum PAC learnability in the same way. Therefore, our undecidability results carry over to quantum PAC learnability as well.

2.2 Gödel Undecidability

For the purpose of this subsection, let F denote a recursively enumerable formal system in which infinitely many different theorems can be proved. (See Definition B.3 for a definition of "recursively enumerable.") Let φ be a primitive recursive enumeration of the theorems provable in F. Here, we think of theorems being "different" in a symbolic way. I.e., two theorems are the same if and only if they are the exact same sequence of symbols from the alphabet available in F. This, in turn, is equivalent to the two theorems having the same Gödel number in a fixed Gödel numbering.

Also, we will denote by $E^2 : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ a primitive recursive enumeration of $\mathbb{N}^2$. I.e., E^2 is a total bijective function such that both component functions $E^2_i : \mathbb{N} \to \mathbb{N}$, $i = 1, 2$, are primitive recursive and such that the inverse $(E^2)^{-1} : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ is primitive recursive. The existence of such an E^2 can, e.g., be proved using so-called pairing functions.

We begin by defining our main object of study for this subsection.

Definition 2.4. *Let F, φ, E^2 be as above. For a compactly supported sequence $a = (a(k))_{k\in\mathbb{N}} \in c_c(\{0,1\})$, define the function $g_a : \mathbb{N} \to \{0,1\}$ via*

$$g_a(n) = \begin{cases} a(n) & \text{if } \varphi(E_1^2(n)) = \neg\varphi(E_2^2(n)) \\ 0 & \text{else} \end{cases},$$

and the function class $\mathcal{G}_F := \{g_a\}_{a\in c_c(\{0,1\})}$.

Here, the equality $\varphi(E_1^2(n)) = \neg\varphi(E_2^2(n))$ is to be understood as the symbolic equality between the theorem with Gödel number $\varphi(E_1^2(n))$ and the negation of the theorem with Gödel number $\varphi(E_2^2(n))$. Equivalently, we require equality of the corresponding Gödel numbers.

We first observe that the class $\mathcal{G}_F$ "collapses" to a single function, the zero function, if and only if the underlying formal system F is consistent.

Proposition 2.5. *F is consistent iff $\mathcal{G}_F = \{0\}$.*

Proof. This follows from the construction of the function class because E_1^2 and E_2^2 are surjective and the range of φ consists exactly of all Gödel numbers of theorems provable in F. □

For later reference, we note a direct consequence of this observation.

Corollary 2.6. *If F is consistent, then* $\mathrm{VCdim}(\mathcal{G}_F) = 0$.

We now make two more observations about the class $\mathcal{G}$. The first concerns its VC-dimension for the case in which the underlying formal system is inconsistent. In that case, the restriction "$\varphi(E_1^2(n)) = \neg\varphi(E_2^2(n))$" is satisfied infinitely often and the VC-dimension of the function class $\mathcal{G}$ is infinite. This is the content of the following

Theorem 2.7. *If F is inconsistent, then* $\mathrm{VCdim}(\mathcal{G}) = \infty$.

For the proof, we first recall that "anything can be deduced from a contradiction," sometimes also known as "ex falso quodlibet."

Proposition 2.8. *Let F be an inconsistent formal system. Let q be a theorem in F. Then both q and $\neg q$ can be proved in F.*

Proof. See Appendix A. □

With this we can now prove Theorem 2.7.

Proof of Theorem 2.7. As F is inconsistent and infinitely many different theorems can be proved in F, by "ex falso quodlibet" there are infinitely many $n \in \mathbb{N}$ such that $\varphi(E_1^2(n)) = \neg\varphi(E_2^2(n))$, because E_1^2 and E_2^2 are surjective and the range of φ consists exactly of all Gödel numbers of theorems provable in F.

Let $N \in \mathbb{N}$. Then, by the above, there exist pairwise distinct $n_1, \dots, n_N \in \mathbb{N}$ such that $\varphi(E_1^2(n_i)) = \neg\varphi(E_2^2(n_i))$ for all $1 \leq i \leq N$. Let $b \in \{0,1\}^N$ be arbitrary. Define $a_b \in c_c(\{0,1\})$ as

$$a_b(n_i) = b_i \text{ for } 1 \leq i \leq N, \quad a_b(n) = 0 \text{ for } n \in \mathbb{N} \setminus \{n_1, \dots, n_N\}.$$

Then we clearly have $g_{a_b}(n_i) = b_i$ for all $1 \leq i \leq N$. So, $\{n_1, \dots, n_N\}$ is shattered by $\mathcal{G}_F$. As $N \in \mathbb{N}$ was arbitrary, we conclude $\mathrm{VCdim}(\mathcal{G}_F) = \infty$. □

If we now combine the statements of Corollary 2.6 and Theorem 2.7, we obtain the following

Corollary 2.9. *F is consistent iff* $\mathrm{VCdim}(\mathcal{G}_F) < \infty$.

Remark 2.10. There is a naïve way of constructing a function class that satisfies the same property as the one just established for $\mathcal{G}_F$. Namely, given F, we could efine

$$\tilde{\mathcal{G}}_F := \begin{cases} \{0\} & \text{if } F \text{ is consistent} \\ c_c(\{0,1\}) & \text{else} \end{cases}.$$

Whereas we can understand $F \mapsto \mathcal{G}_F$ as a computable mapping (see Corollary 2.14), the same is not the case for $F \mapsto \tilde{\mathcal{G}}_F$. Hence, our procedure for constructing $\mathcal{G}_F$ from F has a desirable property that the assignment $F \mapsto \tilde{\mathcal{G}}_F$ would not guarantee.

If the formal system F is capable of expressing both the class $\mathcal{G}_F$ and the finiteness of its VC-dimension, we can combine Corollary 2.9 with Gödel's second incompleteness theorem.

Corollary 2.11. *Assume that F is a recursively enumerable and consistent formal system that contains elementary arithmetic such that infinitely many different theorems can be proved in F. Then* $\mathrm{VCdim}(\mathcal{G}_F) < \infty$*, but the finiteness of* $\mathrm{VCdim}(\mathcal{G}_F)$ *cannot be proved in F.*

Proof. Assume for contradiction that the statement $\mathrm{VCdim}(\mathcal{G}_F) < \infty$ can be proved in F. In Corollary 2.9, we have given a proof that this implies consistency of F. If this proof can be expressed in the formal system F, F proves its own consistency. This contradicts Gödel's second incompleteness theorem. □

Now we come to the second relevant observation about the class $\mathcal{G}_F$: Not only is it a computable function class, but even the mapping $F \mapsto \mathcal{G}_F$ is computable. We first prove the slightly weaker result that $\mathcal{G}_F$ is a computable function class in the sense of Definition C.3:

Theorem 2.12. *Assume that F is a recursively enumerable formal system. Then the class $\mathcal{G}_F$ is computable.*

As a first step towards proving this result, we observe that the sequence space $c_c(\{0,1\})$ used for indexing the class can be recursively enumerated.

Lemma 2.13. *There exists a primitive recursive function* $C : \mathbb{N} \times \mathbb{N} \to \{0,1\}$ *that enumerates* $c_c(\{0,1\})$*, i.e., such that* $c_c(\{0,1\}) = \{n \mapsto C(m,n) \mid m \in \mathbb{N}\}$.

Proof. See Appendix A. □

With this ingredient at hand, we can prove Theorem 2.12.

Proof of Theorem 2.12. According to Definition C.3, we want to find a total computable function $G_F : \mathbb{N} \times \mathbb{N} \to \{0,1\}$ such that $\mathcal{G}_F = \{n \mapsto G_F(m,n) \mid m \in \mathbb{N}\}$. We define

$$G_F(m,n) := \begin{cases} C(m,n) & \text{if } \varphi(E_1^2(n)) = \neg\varphi(E_2^2(n)) \\ 0 & \text{else} \end{cases}.$$

Since C recursively enumerates $c_c(\{0,1\})$, we indeed have $\mathcal{G}_F = \{n \mapsto G_F(m,n) \mid m \in \mathbb{N}\}$. It remains to show that G_F is a total computable function. As C is total computable, even primitive recursive by Lemma 2.13, it suffices to show that the predicate $\varphi(E_1^2(n)) = \neg\varphi(E_2^2(n))$ is total computable.

To this end, recall that E_1^2, E_2^2 and φ are primitive recursive. Thus, we only have to show that, given the Gödel numbers of two theorems, checking whether the theorem corresponding to the first number is the negation of the theorem corresponding to the second number can be done in a computable manner. This is even possible in a primitive recursive manner simply by how Gödel numbers are constructed. □

Note that our proof of Theorem 2.12 even shows that $\mathcal{G}_F$ is primitive recursive if we define a primitive recursive class of functions analogously to Definition C.3. The proof tells us more about the construction of $\mathcal{G}_F$ with respect to computability. Not only is the function class $\mathcal{G}_F$ computable for every formal system F. (This is also true for $\tilde{\mathcal{G}}_F$.) But we even see that the assignment $F \mapsto \mathcal{G}_F$ is computable in the following sense:

Corollary 2.14. *There exists a partial computable function* $\mathbb{G} : \mathbb{N}^3 \to \mathbb{N}$ *such that* $\mathcal{G}_F = \{\mathbb{N} \ni n \mapsto \mathbb{G}(\varphi, m, n) \mid m \in \mathbb{N}\}$ *for any recursively enumerable formal system* F *whose theorems are enumerated by the primitive recursive function* $\varphi : \mathbb{N} \to \mathbb{N}$.

Proof sketch. As φ is primitive recursive, it is in particular computable. Thus, we can represent it via its code with respect to our universal Turing machine. With this code, we can compute the predicate $\varphi(E_1^2(n)) = \neg\varphi(E_2^2(n))$ and the Corollary is proved just like Theorem 2.12. □

Here, the "partial" is only with respect to the first argument, $\mathbb{G}$ is total with respect to the second and third input. Together, Theorem 2.12 and Corollary 2.14 provide an advantage of our construction over the "trivial" $\tilde{G}_F$ in Remark 2.10. Given a recursively enumerable system in terms of an explicit primitive recursive enumeration φ of theorems, they provide us with an explicit algorithmic procedure for evaluating elements of the function class $\mathcal{G}_F$ and thereby with an explicit description of $\mathcal{G}_F$ obtained by fixing certain inputs of the concrete function $\mathbb{G}$.

2.3 Turing Undecidability

We now change the perspective and ask whether there is a general-purpose algorithmic procedure for deciding whether a binary-valued function class has finite VC-dimension. We begin by describing what such a hypothetical algorithm should do: It would take as input the code of an arbitrary computable binary-valued function class $\mathcal{G}$. It should output 0 if $\mathrm{VCdim}(\mathcal{G})$ is infinite and 1 if $\mathrm{VCdim}(\mathcal{G})$ is finite. Note that such an algorithm would decide finiteness of the VC-dimension "only" for computable function classes since it is exactly the computability which allows us to provide their code as input.

We show that such an algorithm does not exist by reduction to the halting problem. The "encoding" of the halting problem into the finiteness of the VC-dimension of a function class is achieved by the following construction.

Definition 2.15. *Let M be a finite-state Turing machine. For a compactly supported sequence $a = (a(k))_{k\in\mathbb{N}} \in c_c(\{0,1\})$, define the function $h_a : \mathbb{N} \to \{0,1\}$ via*

$$h_a(n) = \begin{cases} a(n) & \text{if } M \text{ does not halt after } \leq n \text{ steps on the empty input} \\ 0 & \text{else} \end{cases},$$

and the function class $\mathcal{H}_M := \{h_a\}_{a\in c_c(\{0,1\})}$.

From this definition, we immediately see that whether $\mathrm{VCdim}(\mathcal{H}_M)$ is finite or infinite is determined by whether the underlying Turing machine M halts on the empty input or not.

Lemma 2.16. *Let M be a Turing machine. The binary-valued function class $\mathcal{H}_M$ satisfies*

$$\mathrm{VCdim}(\mathcal{H}_M) = \begin{cases} K & \text{if } M \text{ halts after exactly } K \text{ steps on the empty input} \\ \infty & \text{else} \end{cases}.$$

Proof. First suppose that M halts after exactly $K \in \mathbb{N}_{>0}$ steps on the empty input. Then the set $\{0,\dots,K-1\} \subset \mathbb{N}$ is shattered by $\mathcal{H}_M$. Namely, if $b : \{0,\dots,K-1\} \to \{0,1\}$, then we can append zeros to b to define $a_b \in c_c(\{0,1\})$ via

$$a_b(n) = \begin{cases} b(n) & \text{if } n \leq K-1 \\ 0 & \text{else} \end{cases}, \text{ for } n \in \mathbb{N}.$$

Clearly $h_{a_b}(k) = b(k)$ for all $k \in \{0,\dots,K-1\}$. So $\mathrm{VCdim}(\mathcal{H}_M) \geq K$. As $h_a(n) = 0$ for all $n \geq K$ and for all $a \in c_c(\{0,1\})$, no set of cardinality $\geq K+1$ is shattered by $\mathcal{H}_M$. Thus, also $\mathrm{VCdim}(\mathcal{H}_M) \leq K$.

Now suppose that M does not halt on the empty input. Then, using the same reasoning that gave us the VC-dimension lower bound above, we see that the set $\{0,\dots,N\}$ is shattered by $\mathcal{H}_M$ for every $N \in \mathbb{N}$. Hence, $\mathrm{VCdim}(\mathcal{H}_M) = \infty$. □

Remark 2.17. As in Subsection 2.2, there is a naïve way of constructing a function class with property just established for $\mathcal{H}_M$. Namely, for a Turing machine M, we could define

$$\tilde{\mathcal{H}}_M := \begin{cases} \{0,1\}^{\{0,\dots,K-1\}} & \text{if } M \text{ halts after exactly } K \text{ steps on the empty input} \\ c_c(\{0,1\}) & \text{else} \end{cases},$$

where we think of $\{0,1\}^{\{0,\dots,K-1\}}$ as being embedded into $\{0,1\}^{\mathbb{N}}$ as the first K sequence elements, to which we append zeros. Actually, we have $\mathcal{H}_M = \tilde{\mathcal{H}}_M$, the two function classes are equal. But whereas it might not be obvious from the definition of $\tilde{\mathcal{H}}_M$ that the mapping $M \mapsto \tilde{\mathcal{H}}_M$ is computable, based on Lemma 2.13 it is relatively easy to prove computability of $M \mapsto \mathcal{H}_M$ (see Corollary 2.19). This is why we start from the possibly less intuitive definition of $\mathcal{H}_M = \tilde{\mathcal{H}}_M$.

To use Lemma 2.16 for a reduction to the halting problem, we need to establish two claims. First, we need to show that $\mathcal{H}_M$ is computable according to Definition C.3, so that is makes sense to talk about $\mathcal{H}_M$ as input to a hypothetical algorithm that decides finiteness of the VC-dimension. Only then will $\mathcal{H}_M$, or more precisely the corresponding function H_M, have a code that we can use as input for our hypothetical decision algorithm. Second, we need to show that constructing the class $\mathcal{H}_M$ from the Turing machine M can be done in a computable way. I.e., we need to prove that the mapping $M \mapsto \mathcal{H}_M$ is computable. We begin by establishing computability of $\mathcal{H}_M$.

Theorem 2.18. *Let M be a Turing machine. The function class $\mathcal{H}_M$ is computable.*

Proof. We have already seen in Lemma 2.13 that there exists a primitive recursive function $C : \mathbb{N} \times \mathbb{N} \to \{0,1\}$such that $c_c(\{0,1\}) = \{m \mapsto C(m,n) \mid m \in \mathbb{N}\}$. Therefore, if we define

$$H_M(m,n) = \begin{cases} C(m,n) & \text{if } M \text{ does not halt after } \leq n \text{ steps on the empty input} \\ 0 & \text{else} \end{cases},$$

then $\mathcal{H}_M = \{n \mapsto H_M(m,n) \mid m \in \mathbb{N}\}$. Moreover, H_M is a computable function because is defined from computable functions and a case distinction with a computable predicate. Hence, $\mathcal{H}_M$ is a computable function class according to Definition C.3. □

Computability of $\mathcal{H}_M$ can be seen more easily: $\mathcal{H}_M$ is either finite and thus trivially computable or it is equal to $c_c(\{0,1\})$ and thus computable by Lemma 2.13. We present the proof above because, similarly to our reasoning in Subsection 2.2, it already gives us the computability of $M \mapsto \mathcal{H}_M$:

Corollary 2.19. *There exists a partial computable function $\mathbb{H} : \mathbb{N}^3 \to \mathbb{N}$ such that $\mathcal{H}_M = \{\mathbb{N} \ni n \mapsto \mathbb{H}(M,m,n) \mid m \in \mathbb{N}\}$ for any Turing machine M.*

Again, $\mathbb{H}$ is total with respect to the second and third input. The computability of $M \mapsto \mathcal{H}_M$ is crucial for the final step in our proof of Turing undecidability. And it provides an explicit description of the class $\mathcal{H}_M$ obtained by fixing "input parameters" of the function $\mathbb{H}$.

Now we have everything we need to finish the reduction to the halting problem and thereby our proof of Turing undecidability.

Corollary 2.20. *There is no Turing machine that, upon input of the code of an arbitrary computable binary-valued function class, decides whether that class has finite VC-dimension. In other words, finiteness of the VC-dimension is Turing undecidable.*

Proof. Assume for contradiction that there is such a Turing machine M_{VC}. Then we could construct a Turing machine for solving the halting problem on the empty input as follows:

Given as input the code of a Turing machine M, compute the code of the corresponding class $\mathcal{H}_M$, or, more precisely, the function H_M. This step is possible because the code of a concatenation of Turing machines is a primitive recursive function of their respective codes and because the mapping $M \mapsto \mathcal{H}_M$ is computable by Corollary 2.19. Now feed that code to the Turing machine M_{VC}. If it outputs 1 output, "yes, halts," otherwise output "no, doesn't halt."

As the halting problem is Turing undecidable, we have reached a contradiction. Therefore, the assumed Turing machine does not exist. □

Remark 2.21. We can imitate the construction of $\mathcal{H}_M$ for formal systems. Namely, with F and φ as in Subsection 2.2, we can define, for $a \in c_c(\{0,1\})$,

$$\tilde{g}_a(n) = \begin{cases} a(n) & \text{if } \varphi(1), \dots, \varphi(n) \text{ are consistent} \\ 0 & \text{else} \end{cases},$$

and the function class $\tilde{\mathcal{G}}_F := \{\tilde{g}_a\}_{a \in c_c(\{0,1\})}$. Then, $\mathrm{VCdim}(\tilde{\mathcal{G}}_F) < \infty$ if and only if F is inconsistent. And the mapping $F \mapsto \tilde{\mathcal{G}}_F$ is computable. We see that, for a suitable F, the infiniteness of $\mathrm{VCdim}(\tilde{\mathcal{G}}_F)$ is Gödel undecidable.

Remark 2.22. Both $\mathcal{G}_F$ and $\mathcal{H}_M$ have a finite VC-dimension if and only if they consist only of finitely many distinct functions. Hence, our reasoning implies that, unsurprisingly, (in-)finiteness of a function class is in general Gödel/Turing undecidable. However, as there are infinite function classes with finite VC-dimension, undecidability of the (in-)finiteness of function classes does not yet imply undecidability of the (in-)finiteness of the VC-dimension.

Remark 2.23. One could attempt to derive our Turing undecidability result from Rice's theorem. Informally, Rice's theorem states that any non-trivial semantic property of Turing machines is Turing undecidable [Ric53]. The property "M is a (2-input) Turing machine implementing a class of $\{0,1\}$-valued functions on $\mathbb{N}$ that has finite VC-dimension" is non-trivial and semantic, thus it is Turing undecidable. However, whether a Turing machine implements a class of $\{0,1\}$-valued functions on $\mathbb{N}$ according to Definition C.3 is basically equivalent to whether it halts on every input. Thus, we have found a well known Turing undecidable property hiding in the one above, which makes its undecidability less surprising.

Note the contrast to our result: We only require our hypothetical decision algorithm to work on inputs that describe valid computable function classes. In particular, the algorithm can operate under the premise that it will only receive codes as input that describe total computable functions. Thus, our Turing undecidability result cannot be obtained directly from Rice's theorem.

Remark 2.24. There is a standard way, explained, e.g, in Section 2 of [Poo14], of deriving a Gödel undecidability result from a Turing undecidability result. This allows us to derive from Corollary 2.20: For any recursively enumerable formal system F, there exists a class of $\{0,1\}$-valued functions on $\mathbb{N}$ such that neither finiteness nor infiniteness of its VC-dimension can be proved in F.

The advantage of our reasoning in Subsection 2.2 over this result: Corollary 2.11 provides us with a concrete example of a function class for which finiteness of the VC-dimension is Gödel undecidable. In that sense, the relationship between Corollary 2.11 and the Gödel undecidability result just derived from Corollary 2.20 is analogous to the relationship between Gödel's second and Rosser's [Ros36] strengthening of the first incompleteness theorem.

In fact, starting from a Turing undecidability result, one can derive a Gödel undecidabiltiy result akin to the second incompleteness theorem, compare the essays [Obe19; Cub21]. In our case, starting from Corollary 2.20, given a recursively enumerable formal system F, one can explicitly describe a Turing machine M, depending on F, such that neither $\mathrm{VCdim}(\mathcal{H}_M) < \infty$ nor $\mathrm{VCdim}(\mathcal{H}_M) = \infty$ can be proved in F. This $\mathcal{H}_M$ is then a concrete function class for which finiteness of $\mathrm{VCdim}(\mathcal{H}_M)$ is Gödel undecidable in F and thus gives a result comparable to Corollary 2.11. We have presented our results on independence of the axioms of a formal system and on uncomputability separately, so that these parts of the paper can be read independently from one another.

3 Undecidability in Teaching Problems

In this section, we demonstrate that the function classes constructed in Section 2 are useful beyond the scenario of PAC binary classification, namely also for teaching problems.

3.1 Preliminaries: Teaching Problems and the Teaching Dimension

We now turn our attention to a different learning problem. The differences to the PAC model are two-fold. The source of the training data is now a benevolent teacher who knows the function to be learned. And, instead of requiring the learner to approximate the unknown function with high probability, the unknown function must be exactly identified. To help the learner identify the target function, the teacher has to provide a training data set that uniquely characterizes it. The difficulty of the learning/teaching problem is then captured by the worst case size of a smallest such training data set. This is made formal in the following

Definition 3.1 (Teaching sets and the teaching dimension [GK95])**.** *Let $\mathcal{G} \subset \{0,1\}^{\mathcal{X}}$, $g \in \mathcal{G}$. A set $S = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \{0,1\}$, $N \in \mathbb{N} \cup \{\infty\}$, is a* teaching set *for g in $\mathcal{G}$ if $g(x_i) = y_i$ for all*

$(x_i, y_i) \in S$ *and for every* $\tilde{g} \in \mathcal{G} \setminus \{g\}$ *there exists* $(x_j, y_j) \in S$ *such that* $\tilde{g}(x_j) \neq y_j$. *I.e.,* g *is the unique concept in* $\mathcal{G}$ *that is consistent with the labelled data* S.

The teaching dimension *of* $\mathcal{G}$ *is the worst case size of a minimal teaching set, i.e.,*

$$\mathrm{Tdim}(\mathcal{G}) := \sup_{g \in \mathcal{G}} \inf\{|S| \mid S \text{ is a teaching set for } g\}.$$

We consider a learning/teaching problem for a class $\mathcal{G}$ to be solvable if $\mathrm{Tdim}(\mathcal{G}) < \infty$. Note that we will use this notion specifically for $\mathcal{X} = \mathbb{N}$. This is non-standard. Usually, $\mathcal{X}$ is assumed to be finite so that the teaching dimension is automatically finite.

If the teacher and the learner are allowed to make additional assumptions about the respectively other party's strategy, more refined notions of teaching dimensions should be used (see [Zil+11] for an overview). We, however, restrict our attention to the simplest complexity measure for teaching tasks, namely the one in Definition 3.1.

3.2 Gödel Undecidability of the Existence of Finite Teaching Sets

Before coming to the teaching dimension itself, we discuss a different problem in teaching. Namely, we ask whether, given a function that can be taught to a learner by a teacher using finitely many examples, we can always prove that this is the case. The answer will turn out to be no, in general.

For this and the next subsection, we take F and φ as in Subsection 2.2. We consider the class of threshold functions on $\mathbb{N}$ and allow for the possibility of a "threshold at infinity." I.e., we consider

$$\mathcal{F}_{\mathrm{step}} := \{\mathbb{N} \ni n \mapsto \mathrm{sgn}(n-k) \mid k \in \mathbb{N}\} \cup \{0\},$$

where we use the convention $\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$. Note that $\mathcal{F}_{\mathrm{step}}$ consists of computable functions and is a computable class in the sense introduced in Definition C.3.

We consider the function

$$f_F : \mathbb{N} \mapsto \{0,1\}, \quad f_F(n) = \begin{cases} 0 & \text{if } \varphi(1), \ldots, \varphi(n) \text{ are consistent} \\ 1 & \text{else} \end{cases}.$$

Here, $\varphi(1), \ldots, \varphi(n)$ are said to be inconsistent if and only if for some $1 \leq i, j \leq n$ we have $\varphi(i) = \neg\varphi(j)$, and consistent otherwise.

Note that the mapping $F \mapsto f_F$, where we think of F as given via the code of the corresponding φ, is computable. Clearly, $f_F \in \mathcal{F}_{\mathrm{step}}$ for any formal system F. Therefore we can study whether f_F admits a finite teaching set in the class $\mathcal{F}_{\mathrm{step}}$.

Proposition 3.2. f_F *admits a finite teaching set in* $\mathcal{F}_{step}$ *iff* F *is inconsistent.*

Proof. If F is inconsistent, then there exists $k \in \mathbb{N}$ such that $f_F(n) = \operatorname{sgn}(n-k)$ for all $n \in \mathbb{N}$. So f_F is the only element of $\mathcal{F}_{\text{step}}$ that is consistent with the training data set $\{(k-1,0),(k,1)\}$. Thus, we have found a teaching set of size 2 for f_F.

If F is consistent, then $f_F \equiv 0$ is the zero-function. So any finite training data set consistent with f_F is of the form $\{(n_i,0)\}_{i=1}^N$ for $n_i \in \mathbb{N}$, $1 \leq i \leq N$, $N \in \mathbb{N}$. But also the function $\mathbb{N} \ni n \mapsto \operatorname{sgn}(n-k^*)$ with $k^* = \max_{1\leq i\leq N} n_i + 1$ is an element of $\mathcal{F}_{\text{step}}$ that is consistent with such a training data set. So f_F cannot be uniquely identified in $\mathcal{F}_{\text{step}}$ by a finite training data set. I.e., f_F does not have a teaching set of finite size. □

We see that the teaching dimension of $\mathcal{F}_{\text{step}}$ is infinite. The formal system determines which element of $\mathcal{F}_{\text{step}}$ we consider and Proposition 3.2 states that, if F is consistent, this "filters out" precisely the one concept in $\mathcal{F}_{\text{step}}$ that does not have a finite teaching set.

If F is capable of expressing the function f_F, the class $\mathcal{F}_{\text{step}}$, and the (non-)existence of finite teaching sets, we are again in the position to apply Gödel's second incompleteness theorem.

Corollary 3.3. *Assume that F is a recursively enumerable and consistent formal system that contains elementary arithmetic. The function f_F defined above does not have a finite teaching set in $\mathcal{F}_{step}$, but this statement is not provable in F.*

Remark 3.4. We can use a similar construction to establish an analogous Turing undecidability result. Namely, given a Turing machine M, we can define

$$f_M : \mathbb{N} \mapsto \{0,1\}, \quad f_M(n) = \begin{cases} 0 & \text{if } M \text{ does not halt after } \leq n \text{ steps on the empty input} \\ 1 & \text{else} \end{cases}.$$

f_M admits a finite teaching set in $\mathcal{F}_{\text{step}}$ if and only if M halts on the empty input. Hence, as the mapping $M \mapsto f_M$ is computable, we conclude that there cannot be a general-purpose algorithm that, upon input of a computable function class and a function in that class, decides whether the function admits a finite teaching set in the class.

3.3 Gödel Undecidability of Finiteness of the Teaching Dimension

Next, we study $\mathcal{G}_F$ from the perspective of the teaching dimension. For the purpose of this discussion, F, φ and E^2 are again as in Subsection 2.2. Our first observation is that also finiteness of the teaching dimension of $\mathcal{G}_F$ can be related to consistency of underlying formal system.

Proposition 3.5. *F is consistent iff* $\operatorname{Tdim}(\mathcal{G}_F) < \infty$.

Proof. The proof is similar to that of Corollary 2.9. See Appendix A for details. □

The proof of Proposition 3.5 shows that, if F is inconsistent, then in fact no element of $\mathcal{G}_F$ has a finite teaching set. This is different from our result of the previous subsection, where a

single function in $\mathcal{F}_{\text{step}}$ required teaching sets of infinite size and whether this function was the one characterized by F depended on (in-)consistency.

Again, if F can reason about $\mathcal{G}_F$ and the finiteness of its teaching dimension, we can combine Proposition 3.5 with Gödel's second incompleteness theorem:

Corollary 3.6. *Assume that F is a recursively enumerable and consistent formal system that contains elementary arithmetic. Then $\mathcal{G}_F$ has finite teaching dimension, but this statement cannot be proved in F.*

Thus, we have shown that also the teaching dimension captures the contrast between the "collapse" of $\mathcal{G}_F$ in the consistent case and the "richness" of $\mathcal{G}_F$ in the inconsistent case. Therefore, finiteness of the teaching dimension is also Gödel undecidable.

Remark 3.7. If we leave aside questions of computability, we could also consider the following construction: Take $\tilde{\mathcal{F}}_{\text{step}} \subseteq \{0,1\}^{\mathbb{N}}$ to be the class of proper step functions and consider the class $\{f_F\} \cup \tilde{\mathcal{F}}_{\text{step}}$. This class has finite teaching dimension if and only if F is inconsistent.

3.4 Turing Undecidability of Finiteness of the Teaching Dimension

We can also view $\mathcal{H}_M$ through the lens of teacher-learner interactions. As before, the first step in our approach consists in relating whether the underlying Turing machine M halts to whether the teaching dimension of $\mathcal{H}_M$ is finite.

Proposition 3.8. *Let M be a Turing machine. The binary-valued function class $\mathcal{H}_M$ satisfies*

$$\operatorname{Tdim}(\mathcal{H}_M) = \begin{cases} K & \text{if } M \text{ halts after exactly } K \in \mathbb{N} \text{ steps on the empty input} \\ \infty & \text{else} \end{cases}.$$

Proof. This follows from the equality $\mathcal{H}_{\mathrm{M}} = \tilde{\mathcal{H}}_{\mathrm{M}}$ (Remark 2.17). See Appendix A for details. □

As we already know from Corollary 2.19 that $\mathcal{H}_M$ can be computed from the underlying Turing machine M, we can again reduce to the halting problem and obtain

Corollary 3.9. *There is no Turing machine that, upon input of the code of an arbitrary computable binary-valued function class, decides whether that class has finite teaching dimension. In other words, finiteness of the teaching dimension is Turing undecidable.*

4 Undecidability in Online Learning Problems

As a final demonstration of the applicability of our constructions to different learning models, we show that universal and uniform online learning are undecidable in the by now familiar two senses.

4.1 Preliminaries: Online Learning and Littelstone Trees and Dimension

In online learning, we consider a game between two players, a learner L and an adversary A, both of which know the function class $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$. The game consists of infinitely many rounds. Round $t \in \mathbb{N}_{\geq 1}$ consists of three steps: First, A chooses a "question" $x_t \in \mathcal{X}$. Second, L guesses a label $\hat{y}_t \in \{0,1\}$. Third, A reveals the true label $y_t \in \{0,1\}$ to L. Crucially, A must ensure that the sequence of true labels can actually be realized within $\mathcal{G}$. I.e., the produced sequence $((x_t, y_t))_{t=1}^{\infty}$ must be such that, for every $t \in \mathbb{N}_{\geq 1}$, there exists a function $g \in \mathcal{G}$ with $g(x_s) = y_s$ for all $1 \leq s \leq t$. Note: A does not have to pick a fixed $g \in \mathcal{G}$ in advance. Instead A can choose the true labels adaptively, based on the actions of L and A in previous rounds.

The goal of L is to make as few mistakes as possible, where we say that L makes a mistake in round $t \in \mathbb{N}_{\geq 1}$ if $\hat{y}_t \neq y_t$. Conversely, A wants to make the number of mistakes made by L as large as possible. Note that, while we can also interpret teaching problems as two-player games, the role of the second player is quite different. A teacher is seen as benevolent and has the same goal as the learner. In contrast, an adversary's goal is exactly opposite to that of the learner.

We consider two variants of the online learning problem. On the one hand, we work in the scenario of *universal online learning*, recently introduced in [Bou+20]. We say that $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ is *universally online learnable* if there exists an adaptive strategy $\hat{y}_t = \hat{y}_t(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$ for L such that, for any adversary A, L makes only finitely many mistakes in the above game. On the other hand, we also formulate results in the uniform mistake bound model of online learning, which we refer to as *uniform online learning*. We say that $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ is *uniformly online learnable* if there exist a $d \in \mathbb{N}$ and an (adaptive) strategy $\hat{y}_t = \hat{y}_t(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$ for L such that, for any adversary A, L makes at most d mistakes in the above game.

Both whether a class is universally or uniformly online learnable can be understood in terms of so-called *Littlestone trees*.

Definition 4.1 (Littlestone trees [Lit88; Bou+20]). *A set of points $\{x_{\mathbf{v}}\}_{\mathbf{v} \in \{0,1\}^k, 1 \leq k < d} \subset \mathcal{X}$ is a* Littlestone tree of depth $d \leq \infty$ *of a binary-valued function class $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ if, for every $y_1, \ldots, y_d$ and for every $0 \leq n < d$, there exists $g \in \mathcal{G}$ such that $g(x_{y_1 \ldots y_k}) = y_{k+1}$ holds for all $0 \leq k \leq n$. We say that $\mathcal{G}$* has an infinite Littlestone tree *if there exists a Littlestone tree of depth ∞ of $\mathcal{G}$.*

A Littlestone tree of $\mathcal{G}$ is a complete binary tree in which the nodes are labelled by points in $\mathcal{X}$ and the edges are labelled by 0 or 1 in such a way that for every path of finite length, starting from the root of the tree, there is a function in $\mathcal{G}$ that labels all nodes along the path according to the respectively outgoing edges. Note that the definition is only concerned with finite paths, even for an infinite Littlestone tree.

The relation between universal online learnability and the (non-)existence of infinite Littlestone trees is summarized in the following

Theorem 4.2 (Theorem 3.1 in [Bou+20]). *$\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ is universally online learnable iff $\mathcal{G}$ does not have an infinite Littlestone tree.*

Going from "universal" to "uniform" on the level of Littlestone trees corresponds to requiring a uniform bound on the depth of all Littlestone trees of a class. This gives rise to

Definition 4.3 (Littlestone dimension [Lit88])**.** *Let $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$. The Littlestone dimension of $\mathcal{G}$ is defined to be*

$$\mathrm{Ldim}(\mathcal{G}) := \sup\{d \in \mathbb{N}_0 \mid \mathcal{G} \text{ has a Littlestone tree of depth } d\}.$$

Note that, if $\mathcal{G}$ has an infinite Littlestone tree, then $\mathrm{Ldim}(\mathcal{G}) = \infty$. The converse, however, is not true, as $\mathrm{Ldim}(\mathcal{G}) = \infty$ also holds if $\mathcal{G}$ has Littlestone trees of arbitrarily large depth but no infinite Littlestone tree.

The Littlestone dimension characterizes uniform online learnability according to the following

Theorem 4.4 (Theorem 3 in [Lit88])**.** *$\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ is uniformly online learnable with at most $d \in \mathbb{N}$ mistakes iff $\mathrm{Ldim}(\mathcal{G}) \leq d$. In particular, $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ is uniformly online learnable iff $\mathrm{Ldim}(\mathcal{G}) < \infty$.*

4.2 Gödel Undecidability of Finiteness of the Littlestone Dimension

We have seen in Theorem 4.4 that uniform online learnability is equivalent to the Littlestone dimension being finite. Therefore, we again start by relating consistency of the formal system F underlying $\mathcal{G}_F$ to finiteness of the Littlestone dimension of $\mathcal{G}_F$. For both this subsection and for Subsection 4.4, we use the notation and classes introduced in Subsection 2.2.

Proposition 4.5. *F is consistent iff $\mathrm{Ldim}(\mathcal{G}_F) < \infty$.*

Proof. This follows from our results on the VC-dimension and the inequality $\mathrm{VCdim} \leq \mathrm{Ldim}$. See Appendix A for details. □

The Gödel undecidability of uniform online learnability now follows as in Subsection 2.2:

Corollary 4.6. *Assume that F is a recursively enumerable and consistent formal system that contains elementary arithmetic. Then $\mathcal{G}_F$ has finite Littlestone dimension, but this statement cannot be proved in F.*

4.3 Turing Undecidability of Finiteness of the Littlestone Dimension

This subsection as well as Subsection 4.5 use the notation and constructions from Subsection 2.3. With Theorem 4.4 and the results from Subsection 2.3 in place, the only step left is to observe that $\mathcal{H}_M$ has finite Littlestone dimension iff M halts.

Proposition 4.7. *Let M be a Turing machine. The binary-valued function class $\mathcal{H}_M$ satisfies*

$$\mathrm{Ldim}(\mathcal{H}_M) = \begin{cases} K & \text{if } M \text{ halts after exactly } K \text{ steps on the empty input} \\ \infty & \text{else} \end{cases}.$$

Proof. This follows from our results on the VC-dimension and the inequalities $\mathrm{VCdim}(\mathcal{H}_M) \leq \mathrm{Ldim}(\mathcal{H}_M) \leq \log_2|\mathcal{H}_M|$. See Appendix A for details. □

Now, the line of reasoning presented in Subsection 2.3 implies the Turing undecidability of uniform online learnability:

Corollary 4.8. *There is no Turing machine that, upon input of the code of an arbitrary computable binary-valued function class, decides whether that class has finite Littlestone dimension. In other words, finiteness of the Littlestone dimension is Turing undecidable.*

4.4 Gödel Undecidability of the Existence of Infinite Littlestone Trees

Because of Theorem 4.2, we first establish an equivalence between the formal system F underlying the class $\mathcal{G}_F$ being consistent and $\mathcal{G}_F$ having no infinite Littlestone tree.

Proposition 4.9. *F is consistent iff $\mathcal{G}_F$ does not have an infinite Littlestone tree.*

Proof. The proof is similar to that of Corollary 2.9. See Appendix A. □

With this observation, the same reasoning, using Gödel's second incompleteness theorem, as in Subsection 2.2 yields:

Corollary 4.10. *Assume that F is a recursively enumerable and consistent formal system that contains elementary arithmetic. Then $\mathcal{G}_F$ does not have an infinite Littlestone tree, but this statement cannot be proved in F.*

4.5 Turing Undecidability of the Existence of Infinite Littlestone Trees

Analogously to the other scenarios, we want to connect the (non-)existence of infinite Littlestone trees for $\mathcal{H}_M$ to whether or not M halts.

Proposition 4.11. *Let M be a Turing machine. The binary-valued function class $\mathcal{H}_M$ has an infinite Littlestone tree iff M does not halt on the empty input.*

Proof. The proof is similar to that of Lemma 2.16. See Appendix A for details. □

As before, because of the computability of $M \mapsto \mathcal{H}_M$, through Theorem 4.2, this implies that universal online learnability is Turing undecidable:

Corollary 4.12. *There is no Turing machine that, upon input of the code of an arbitrary computable binary-valued function class, decides whether that class has an infinite Littlestone tree. In other words, the existence of infinite Littlestone trees is Turing undecidable.*

5 Conclusion

In this work, we have shown that in the standard model of binary classification, in two models of online learning, and in a basic model describing teacher-learner interactions, it is in general undecidable whether the learning task can be completed. We have established this for two different meanings of "undecidable," the first being "true, but not provable in a formal system" and the second being "not computable." In both cases, our results follow by providing computable constructions that allow for a reduction of the problem of deciding finiteness of the complexity measure for the respective learning task to the prototypic undecidable problem, i.e., to proving consistency of a formal system or to deciding whether a Turing machine halts.

It was already known, due to [Ben+19], that learnability can be independent of the axioms of ZFC. We have proved a similar-in-spirit result for the arguably most influential learning model, the PAC model of binary classification. By discussing our proof strategy also for a teacher-learner model and for online learning, we have demonstrated that it is not specific to the PAC setting. Moreover, learnability can be undecidable also in other formal systems and in the terminology of computer science. A crucial feature of our constructions, especially for establishing undecidability in the latter sense, is that we are only dealing with computable objects. This is to be contrasted with [Ben+19], where the continuum is used. In particular, the arguments of [Ben+19] do not give uncomputability results. Note that, because our constructions are computable, instead of PAC learnability, equivalently we could have considered computable PAC learnability, because of Theorem 10 of [Aga+20], when restricting our attention to the realizable scenario.

We hope that our work adds to the ongoing research aiming towards a better understanding of the theoretical prospects and limits of machine learning. Our results indicate that potential problems for applications of machine learning do not only arise on the level of algorithmic design, which in itself is an extremely challenging task. Rather, already when faced with a task, we encounter a fundamental difficulty: It is in general not possible to decide whether that task is, from the information-theoretic perspective of sample complexity, leaving questions of computational complexity aside, learnable, i.e., in principle amenable to a solution via machine learning.

From a more practical perspective, our results can be interpreted as follows: When faced with a learning task, one can usually choose which hypothesis class to use. This choice will be guided by different considerations, such as prior knowledge about the problem, potential issues for optimization, and questions of learnability. In particular, one usually chooses a class that is known to be learnable. Thereby, the "library" of candidate function classes is restricted to those whose learnability has already been established. Our results say that there is no generic way of enlarging this library: Every time one faces a learning problem for which all classes from the current library perform poorly, identifying a new suitable candidate class, even leaving questions of optimization aside, presents a new challenge because of learnability alone.

Finally, we mention some questions raised by our work:

- We have approached learnability through criteria based on complexity measures of the function class under consideration. Can (un-)decidability be established for learnability via algorithmic properties, e.g., stability or compression-based schemes?
- For our PAC learning scenario, we require a sample complexity bound that is uniform over the function class. Can our results be extended to non-uniform learning models in which the sample size is allowed to depend on the function to be learned, e.g., via some "weight parameter"?
- As discussed in Remark 2.22, it would be interesting to see whether the finiteness of the VC-, teaching or Littlestone dimension remains undecidable also when restricting the potential inputs to codes of infinite function classes.
- As observed in [Bou+20], universal online learning is closely connected to Gale-Stewart games. Do undecidability results in one of these two scenarios translate to the respectively other one? For example, can we recover the undecidability results for Gale-Stewart games due to [Rab58] and [Jon82] from our results in Section 4? Or can we these works to gain further insight into undecidability in online learning?

Acknowledgements

I want to thank Michael M. Wolf for stimulating discussions on questions of (un-)decidability and for suggesting the reasoning used in Subsection 3.2. I also thank the anonymous reviewers and the meta-reviewer from COLT 2021 for their feedback. Furthermore, I thank Artem Chernikov, Asaf Karagila, Aryeh Kontorovich, Vladimir Pestov and, Roi Weiss for pointing the measure-theoretic subtleties around Theorem 2.3 out to me. Finally, I thank Aryeh Kontorovich for bringing the references [PY96; MU02] and [Han+21] to my attention.

Support from the TopMath Graduate Center of TUM the Graduate School at the Technische Universität München, Germany, from the TopMath Program at the Elite Network of Bavaria, and from the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes) is gratefully acknowledged.

References

[Aga+20] S. Agarwal, N. Ananthakrishnan, S. Ben-David, T. Lechner, and R. Urner. "On Learnability wih Computable Learners". In: *Algorithmic Learning Theory* (2020), pp. 48–60. ISSN: 1938-7228. URL: http://proceedings.mlr.press/v117/agarwal20b.html (cited on pp. 5, 21).

[AW18] Srinivasan Arunachalam and Ronald de Wolf. "Optimal quantum sample complexity of learning algorithms". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2879–2878. DOI: 10.5555/3291125.3309633 (cited on p. 7).

[Bar93] J. Barwise, ed. *Handbook of mathematical logic*. 8. impr. Vol. 90. Studies in logic and the foundations of mathematics. Amsterdam [u.a.]: North-Holland Publ, 1993. ISBN: 978-0444863881 (cited on p. 29).

[Ben+19] S. Ben-David, P. Hrubeš, S. Moran, A. Shpilka, and A. Yehudayoff. "Learnability can be undecidable". In: *Nature Machine Intelligence* 1.1 (2019), pp. 44–48. ISSN: 2522-5839. DOI: 10.1038/s42256-018-0002-3 (cited on pp. 5, 21).

[Blu+89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. "Learnability and the Vapnik-Chervonenkis dimension". In: *Journal of the ACM (JACM)* 36.4 (1989), pp. 929–965 (cited on p. 7).

[Bou+20] O. Bousquet, S. Hanneke, S. Moran, R. van Handel, and A. Yehudayoff. *A Theory of Universal Learning*. Version 1. Nov. 9, 2020. arXiv: 2011.04483 [cs.LG] (cited on pp. 4, 18, 22).

[Cub21] T. S. Cubitt. *A Note on the Second Spectral Gap Incompleteness Theorem*. Version 1. May 20, 2021. arXiv: 2105.09854 [quant-ph] (cited on p. 14).

[Dav82] M. Davis. *Computability and unsolvability*. Dover books on advanced mathematics. New York: Dover, 1982. ISBN: 978-0486614717 (cited on p. 29).

[Dud78] R. M. Dudley. "Central limit theorems for empirical measures". In: *The Annals of Probability* (1978), pp. 899–929 (cited on p. 7).

[End13] H. B. Enderton. *A mathematical introduction to logic*. 3rd ed. Oxford: Academic, 2013. ISBN: 978-0123869777 (cited on p. 29).

[FL98] M. Frances and A. Litman. "Optimal mistake bound learning is hard". In: *Information and Computation* 144.1 (1998), pp. 66–82. DOI: 10.1006/inco.1998.2709 (cited on p. 6).

[Gan20] A. Gandolfi. *Decidability of Sample Complexity of PAC Learning in finite setting*. Version 1. Feb. 26, 2020. arXiv: 2002.11519 [cs.LG] (cited on p. 5).

[GK95] S. A. Goldman and M. J. Kearns. "On the Complexity of Teaching". In: *Journal of Computer and System Sciences* 50.1 (1995), pp. 20–31. ISSN: 00220000. DOI: 10.1006/jcss.1995.1003 (cited on pp. 4, 14).

[Göd31] K. Gödel. "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". In: *Monatshefte für Mathematik und Physik* 38.1 (1931), pp. 173–198. ISSN: 1436-5081. DOI: 10.1007/BF01700692. URL: https://link.springer.com/article/10.100 (cited on p. 29).

[Han+21] S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. *Universal Bayes consistency in metric spaces*. Version 7. Jan. 6, 2021. arXiv: 1906.09855 [cs.LG] (cited on pp. 5, 6, 23).

[Har19] K. P. Hart. "Machine learning and the continuum hypothesis". English. In: *Nieuw Arch. Wiskd. (5)* 20.3 (2019), pp. 214–217. ISSN: 0028-9825 (cited on p. 5).

[Jon82] J. P. Jones. "Some undecidable determined games". In: *International Journal of Game Theory* 11.2 (1982), pp. 63–70. ISSN: 0020-7276. DOI: 10.1007/BF01769063 (cited on p. 22).

[Kle02] S. C. Kleene. *Mathematical logic*. Dover ed. Mineola, N.Y.: Dover Publications, 2002. ISBN: 0486425339 (cited on p. 29).

[Las92] M. C. Laskowski. "Vapnik-Chervonenkis classes of definable sets". In: *Journal of The London Mathematical Society-second Series* 45 (1992), pp. 377–384 (cited on p. 5).

[Lat96] R. H. Lathrop. "On the learnability of the uncomputable". In: *Proc. 13th International Conference on Machine Learning*. Morgan Kaufmann, 1996, pp. 302–309 (cited on p. 4).

[Lit88] N. Littlestone. "Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm". In: *Machine Learning* 2.4 (1988), pp. 285–318. ISSN: 1573-0565. DOI: 10.1023/A:1022869011914 (cited on pp. 4, 18, 19).

[LMR91] N. Linial, Y. Mansour, and R. L. Rivest. "Results on learnability and the Vapnik-Chervonenkis dimension". In: *Information and Computation* 90.1 (1991), pp. 33–49. DOI: 10.1016/0890-5401(91)90058-A (cited on p. 6).

[MR17] P. Manurangsi and A. Rubinstein. "Inapproximability of VC Dimension and Littlestone's Dimension". In: *Proceedings of the 2017 Conference on Learning Theory*. Ed. by S. Kale and O. Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, July 2017, pp. 1432–1460. URL: http://proceedings.mlr.press/v65/manurangsi17a.html (cited on p. 6).

[MU02] E. Mossel and C. Umans. "On the complexity of approximating the VC dimension". In: *Journal of Computer and System Sciences* 65.4 (2002), pp. 660–671. DOI: 10.1016/S0022-0000(02)000 (cited on pp. 6, 23).

[Obe19] S. Oberhoff. *Incompleteness Ex Machina*. Version 1. Sept. 6, 2019. arXiv: 1909.04569 [cs.LO] (cited on p. 14).

[Pes11] V. Pestov. "PAC learnability versus VC dimension: a footnote to a basic result of statistical learning". In: *The 2011 International Joint Conference on Neural Networks*. IEEE. 2011, pp. 1141–1145 (cited on p. 7).

[Pol84] D. Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. New York, NY: Springer, 1984. ISBN: 9781461297581 (cited on p. 7).

[Poo14] B. Poonen. "Undecidable problems: A sampler". In: *Interpreting Gödel: Critical Essays*. Cambridge University Press, 2014, pp. 211–241. DOI: 10.1017/CBO9780511756306.015 (cited on p. 14).

[PY96] C. H. Papadimitriou and M. Yannakakis. "On limited nondeterminism and the complexity of the VC dimension". In: *Journal of Computer and System Sciences* 53.2 (1996), pp. 161–170. DOI: 10.1006/jcss.1996.0058 (cited on pp. 6, 23).

[Rab58] M. O. Rabin. "Effective computability of winning strategies". In: *Contributions to the Theory of Games (AM-39), Volume III*. Ed. by M. Dresher, A. W. Tucker, and P. Wolfe. Princeton University Press, 1958, pp. 147–158. DOI: doi:10.1515/9781400882151-008 (cited on p. 22).

[Ric53] H. G. Rice. "Classes of recursively enumerable sets and their decision problems". In: *Transactions of the American Mathematical Society* 74.2 (1953), pp. 358–366 (cited on p. 13).

[Ros36] B. Rosser. "Extensions of some theorems of Gödel and Church". In: *Journal of Symbolic Logic* 1.3 (1936), pp. 87–91. ISSN: 0022-4812. DOI: 10.2307/2269028 (cited on p. 14).

[SB19] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. 12th printing. Cambridge: Cambridge University Press, 2019. ISBN: 978-1107057135 (cited on p. 7).

[Sch00] M. Schaefer. "Deciding the K-dimension is PSPACE-complete". In: *Proceedings 15th Annual IEEE Conference on Computational Complexity*. IEEE. 2000, pp. 198–203. DOI: 10.1109/CCC.2000.856750 (cited on p. 6).

[Sch99] M. Schaefer. "Deciding the Vapnik–Červonenkis Dimension is Σ_3^p-complete". In: *Journal of Computer and System Sciences* 58.1 (1999), pp. 177–182. DOI: 10.1006/jcss.1998.1602 (cited on pp. 5, 6).

[SFM21] S. Sehra, D. Flores, and G. D. Montanez. *Undecidability of Underfitting in Learning Algorithms*. Version 3. Feb. 9, 2021. arXiv: 2102.02850 [cs.ML] (cited on p. 5).

[Shi93] A. Shinohara. "Complexity of computing Vapnik-Chervonenkis dimension". In: *International Workshop on Algorithmic Learning Theory*. Springer. 1993, pp. 279–287. DOI: 10.1007/3-540-57370-4_54 (cited on p. 6).

[Soa16] R. I. Soare. *Turing Computability*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. ISBN: 978-3-642-31932-7. DOI: 10.1007/978-3-642-31933-4 (cited on p. 29).

[Soa78] Robert I Soare. "Recursively enumerable sets and degrees". In: *Bulletin of the American Mathematical Society* 84.6 (1978), pp. 1149–1181 (cited on p. 5).

[Sol08] D. Soloveichik. *Statistical learning of arbitrary computable classifiers*. Version 2. July 10, 2008. arXiv: 0806.3537 [cs.LG] (cited on p. 5).

[Tay19] W. Taylor. *Learnability can be independent of zfc axioms: Explanations and implications*. Version 1. Sept. 16, 2019. arXiv: 1909.08410 [cs.LG] (cited on p. 5).

[Tur37] A. M. Turing. "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* s2-42.1 (1937), pp. 230–265. ISSN: 0024-6115. DOI: 10.1112/plms/s2-42.1.230 (cited on pp. 29, 30).

[Val84] L. G. Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142. ISSN: 00010782. DOI: 10.1145/1968.1972 (cited on p. 6).

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability & Its Applications* 16.2 (1971), pp. 264–280. DOI: 10.1137/1116025 (cited on pp. 2, 7).

[Zha18] K. Zhao. *A statistical learning approach to a problem of induction*. Dec. 8, 2018. URL: http://philsci-archive.pitt.edu/15422/ (cited on p. 5).

[Zil+11] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. "Models of Cooperative Teaching and Learning". In: *J. Mach. Learn. Res.* 12 (2011), pp. 349–384. URL: http://portal.acm.org/citation.c (cited on p. 15).

Appendix

A Proofs

Proof of Proposition 2.8. As F is inconsistent, there exists a theorem p such that both p and $\neg p$ can be proved in F. As p can be proved in F, we have $q \to p$ (which is our notation for "q implies p"). By negation we then have $\neg p \to \neg q$. As $\neg p$ can be proved in F, also $\neg q$ can be proved in F. If we now exchange q by $\neg q$ in the above reasoning, we see that also q can be proved in F. □

Proof of Lemma 2.13. We define $C : \mathbb{N} \times \mathbb{N} \to \{0, 1\}$ via

$$C(m, n) = \begin{cases} n^{\text{th}} \text{ bit in the binary representation of } m & \text{if } m > 0 \text{ and } 2^n \leq m \\ 0 & \text{else} \end{cases}.$$

As exponentiation and finding the binary representation of a natural number can be done in a primitive recursive manner, the function C is defined in terms of primitive recursive functions and a case distinction via a primitive recursive predicate Thus, C is itself primitive recursive.

Clearly, the function $n \mapsto C(m, n)$ has finite support and is thus an element of $c_c(\{0, 1\})$. Conversely, if $a \in c_c(\{0, 1\})$, then there exists $K \in \mathbb{N}$ such that $a(n) = 0$ for all $n > K$. Hence, if we take m_a to be the natural number with binary representation $a(0) \dots a(K)$, then $a(n) = C(m_a, n)$ for all $n \in \mathbb{N}$. □

Proof of Proposition 3.5. If F is consistent, $\mathcal{G}_F = \{0\}$ and the claim is trivial. If F is inconsistent, then uniquely identifying a function $g_a \in \mathcal{G}_F$ requires one to uniquely identify the subsequence $(a_{k_l})_{l \in \mathbb{N}}$ of $a \in c_c(\{0, 1\})$ chosen such that $k_{l+1} > k_l$ and such that $\varphi(E_1^2(k)) = \neg\varphi(E_2^2(k))$ iff $k = k_l$ for some $l \in \mathbb{N}$. As $\varphi(E_1^2(k)) = \neg\varphi(E_2^2(k))$ is satisfied for infinitely many $k \in \mathbb{N}$ (see Proposition 2.8) and the size of the support of an element of $c_c(\{0, 1\})$ can be arbitrarily large, any training data set that uniquely identifies g_a has to consist of infinitely many labelled examples. □

Proof of Proposition 4.5. If F is consistent, $\mathcal{G}_F = \{0\}$ and clearly $\text{Ldim}(\mathcal{G}_F) = 0$. If F is inconsistent, we can use the well known inequality VCdim $\leq$ Ldim together with Theorem 2.7 to obtain $\text{Ldim}(\mathcal{G}_F) = \infty$. □

Proof of Proposition 4.7. This follows quite directly the well known fact that, for any function class $\mathcal{G} \subset \{0, 1\}^{\mathcal{X}}$, $\text{VCdim}(\mathcal{G}) \leq \text{Ldim}(\mathcal{G}) \leq \log_2|\mathcal{G}|$.

Namely, if M halts on the empty input, these two inequalities, due to Lemma 2.16 and Remark 2.17, become $K \leq \text{Ldim}(\mathcal{G}) \leq \log_2|\mathcal{H}| = K$. And if M does not halt on the empty input, the lower bound via the VC-dimension, together with Remark 2.17, implies $\text{Ldim}(\mathcal{G}) = \infty$. □

Proof of Proposition 4.9. If F is consistent, then $\text{Ldim}(\mathcal{G}_F) < \infty$ by Proposition 4.5. In particular, $\mathcal{G}_F$ does not have an infinite Littlestone tree.

If F is inconsistent, then, as we have seen in the proof of Theorem 2.7, there exists a sequence $(n_k)_{k=1}^{\infty} \subset \mathbb{N}$ such that $\{n_1, \dots, n_N\}$ is shattered by $\mathcal{G}_F$ for every $N \in \mathbb{N}_{\geq 1}$. Therefore, we obtain an infinite Littlestone tree of $\mathcal{H}_M$ by labelling every node in the k^{th} layer by n_{k+1}, for $k \in \mathbb{N}_0$. □

Proof of Proposition 4.11. If M halts on the empty input, then $\operatorname{Ldim}(\mathcal{H}_M) < \infty$ by Proposition 4.7. In particular, $\mathcal{H}_M$ does not have an infinite Littlestone tree.

If M does not halt on the empty input, then $\{0, \dots, N\}$ is shattered by $\mathcal{H}_M$ for every $N \in \mathbb{N}$, as we have seen in the proof of Lemma 2.16. Therefore, we obtain an infinite Littlestone tree of $\mathcal{H}_M$ by labelling every node in the k^{th} layer by k, for $k \in \mathbb{N}_0$. □

B Gödel and Incompleteness of Formal Systems

Here, we compile standard notions connected to formal systems which appear in the main body of the paper. However, some notions will only be introduced informally and the interested reader is referred to other sources for the formal definitions.

We denote by $\mathbb{N}$ the natural numbers including 0. We call a function $f : \mathbb{N}^k \to \mathbb{N}$ *primitive recursive* if it can be built from the zero function, the successor function, and the coordinate projection functions via composition and primitive recursion. From a modern perspective, the primitive recursive functions are those that can be implemented using basic arithmetic as well as IF THEN ELSE, AND, OR, NOT, $=$, $>$, and FOR loops. WHILE loops are not allowed here.

Next, we recall, albeit only informally, the notion of a formal system.

Definition B.1 (Formal systems - Informal)**.** *A* formal system F *consists of a finite alphabet of symbols, a language of statements that can be well-formed from the alphabet, a distinguished set of statements called* axioms*, and rules for how to derive/prove new theorems from these axioms.*
A formal system F is called consistent *if there is no well-formed statement such that both it and its negation can be proved in F. Otherwise, we call F* inconsistent*.*

We will be interested in a particular kind of formal systems in which the provable theorems, i.e., the statements that can be deduced from the axioms according to the derivation rules, can be recursively enumerated. To make this assumption more rigorous, we first recall

Definition B.2 (Gödel numbering - Informal)**.** *A* Gödel numbering *for a formal system F is an injective function that maps each symbol in the alphabet and each well-formed statement to an element of* $\mathbb{N}$*.*

For our purposes, it does not matter which Gödel numbering is used. We only use that Gödel numberings exist for which "translating" between a string of Gödel numbers of symbols describing a statement and the actual Gödel number of that statement can be done primitive recursively in both directions. Gödel's original construction has this property. From now on, we fix such a Gödel numbering. This allows us to identify statements in a formal system with elements of $\mathbb{N}$ and

"manipulations" of statements with primitive recursive maps between natural numbers. Both of these identifications will sometimes be implicit throughout the paper.

From this perspective, we can describe the type of formal systems used in this work.

Definition B.3 (Recursively enumerable formal systems)**.** *A formal system F is called* recursively enumerable *(or* effectively axiomatized*) if there exists a primitive recursive function $\varphi : \mathbb{N} \to \mathbb{N}$ such that $\{\varphi(n) \mid n \in \mathbb{N}\}$ is exactly the set of all Gödel numbers in a fixed Gödel numbering of statements that can be proved in F.*

Given such a primitive recursive enumeration φ of provable theorems, we will sometimes abuse notation and take $\varphi(n)$ to denote both a theorem and its Gödel number. The exact meaning, if not made explicit, will be clear from the context.

Gödel's second incompleteness theorem provides, for any recursively enumerable and consistent formal system that contains elementary arithmetic, an explicit statement that is true but cannot be proved in that formal system.

Theorem B.4 (Gödel's second incompleteness theorem [Göd31])**.** *Assume that F is a recursively enumerable and consistent formal system that contains elementary arithmetic. Then the consistency of F is not provable in F.*

We call a statement that is true but not provable in a formal system F *Gödel undecidable* in F. This is not standard terminology, we merely use it to shorten some formulations.

For a more formal presentation of these and other notions from mathematical logic, the reader is referred to textbooks such as [Bar93; Kle02; End13].

C Turing and Uncomputability

This section recalls standard definitions and results related to Turing machines and computability. Again, sometimes we give only an informal presentation and refer to textbooks for details.

In [Tur37], Turing introduced what are now known as a Turing machines. We do not give a formal definition, but instead describe the workings of a Turing machine informally. For a more rigorous presentation, see, e.g., [Dav82; Soa16].

Definition C.1 (Turing machines - Informal)**.** *A Turing machine M consists of*

- *a 1-dimensional tape with infinitely many cells extending in both directions, each of which contains a symbol from a finite alphabet Σ,*
- *a head that can read and write symbols in a single cell and move to the left or to the right by one cell,*
- *a finite set of states Q containing an* initial state *and a* halting state*,*

- *and an instruction function* $I : \Sigma \times Q \to \Sigma \times \{L, R\} \times Q$ *describing the write-, move- and state-update-behaviour of* M *upon reading a given symbol while in a given state.*

The two distinguished states are the initial state, in which the Turing machine begins any of its computations, and the halting state, that causes the Turing machine to halt when it is reached.

According to the Church-Turing thesis, which could be considered a "law of nature" for the world of computing, everything that can be reasonably considered computable is computable by a Turing machine. Hence, we take Turing machines as our model for defining computability.

Definition C.2 ((Turing) Computable functions)**.** *A partial function* $f : \mathbb{N}^k \to \mathbb{N}$*, for* $k \in \mathbb{N}_{\geq 1}$*, is* (Turing) computable *if there exists a finite-state Turing machine* M *such that, whenever we run* M *on a tape with an encoding of* $x \in \mathrm{dom}(f)$ *written on it,* M *eventually halts with the tape containing an encoding of* $f(x)$*, and whenever we run* M *on a tape with an encoding of* $x \notin \mathrm{dom}(f)$ *written on it,* M *does not halt.*

One possible choice of encoding is the unary encoding. I.e., $x \in \mathbb{N}$ is represented by $x + 1$ consecutive ones on the tape. The remaining tape is left blank. An element of $\mathbb{N}^k$ can be represented by k blocks of unary encodings of the components, separated by single zeros.

It is useful to note at this point that any primitive recursive function is computable. However, there are computable functions that are not primitive recursive.

We call a decision problem whose corresponding function, mapping instances of the problem to a binary "yes-or-no" output, is not computable *Turing undecidable*. The prototypic example of a Turing undecidable decision problem is the *halting problem*, i.e., the problem of deciding whether a given Turing machine halts on the empty input. Already [Tur37] observed that this cannot be achieved in a computable way.

We will also use a notion of computability of function classes.

Definition C.3. *We say that a class* $\mathcal{G} \subseteq \mathbb{N}^{\mathbb{N}}$ *is computable if there exists a total computable function* $G : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ *such that* $\mathcal{G} = \{n \mapsto G(m, n) \mid m \in \mathbb{N}\}$*.*

We recall one last fact related to Turing machines. Namely, there exist *universal Turing machines* capable of simulating any Turing machine [Tur37]. From now on, for each $k \in \mathbb{N}_{\geq 1}$, we fix such a universal Turing machine understood as a partial computable function $\mathbb{M} : \mathbb{N}^{k+1} \to \mathbb{N}$. Then, for any Turing machine M and corresponding partial computable function $f_M : \mathbb{N}^k \to \mathbb{N}$, there exists a natural number, also denoted by M, such that $f_M(x) = \mathbb{M}(M, x)$ for every $x \in \mathbb{N}$. The natural number M is called the *code* of the Turing machine M with respect to $\mathbb{M}$. This allows us to think of Turing machines, or, equivalently, computable functions, as input when representing them by their code with respect to our fixed universal Turing machine.

Appendix D

Articles as co-author

D.1 Quantum and classical dynamical semigroups of superchannels and semicausal channels

Quantum and classical dynamical semigroups of superchannels and semicausal channels

Markus Hasenöhrl and Matthias C. Caro

Quantum superchannels describe transformations from quantum channels to quantum channels. In particular, a quantum circuit board with a free slot, which a quantum channel can be plugged in, constitutes a physical realization of such a quantum superchannel. Actual implementations of such quantum circuit boards will be subject to decay, leading to a change of the corresponding superchannel over time. For many cases of interest, the resulting family of superchannels for different times will be Markovian and form a continuous one-parameter semigroup. In this work, we completely characterize the generators giving rise to such semigroups of superchannels, both in the classical and the quantum case.

We begin this work in Section I with an example motivating our focus on continuous one-parameter semigroups of quantum superchannels, which we refer to as quantum dynamical semigroups of superchannels. The introductory Section also contains an overview over the structure of the paper and a short review of related work. This is followed by Section II, in which we present and discuss informal versions of our results. Section III contains mathematical preliminaries, in preparation for the remainder of the paper.

In Section IV, we prove our results for continuous one-parameter semigroups of classical superchannels. After introducing the classical analogues of superchannels and semicausal channels, we state and prove two classical analogues of known results from quantum information theory in Subsections IV.A and IV.B: First, classical superchannels correspond to certain semicausal nonnegative maps (Theorem IV.3). And second, classical semicausality is equivalent to semilocalizability (Theorem IV.4). We exploit the latter equivalence in Subsection IV.C to establish a normal form for all possible generators of continuous one-parameter semigroups of semicausal nonnegative maps (Theorem IV.7). Namely, we identify two basic building blocks such that all admissible generators can be obtained as convex combinations thereof. The results of Subsection IV.A now allow us to translate this to a normal form for the generators continuous one-parameter semigroups of classical superchannels in Subsection IV.D (Theorem IV.10). For our proofs in Section IV, we use vectorization as a classical version of the Choi-Jamiolkowski isomorphism, the known characterization of generators of general semigroups of nonnegative matrices, and linear-algebraic techniques.

Armed with the intuition from the classical case, we turn to the quantum setting in Section V. The overall line of reasoning here mirrors that of Section IV. First, in Subsection V.A, we recall the known correspondence between quantum superchannels and certain quantum operations (Theorem V.3), and then extend the equivalence between quantum semicausality and quantum semilocalizability to infinite dimensions (Theorem V.4). Subsection V.B now contains our main technical contributions: We give an efficiently checkable criterion for a linear map to be a valid generator giving rise to a semigroup of semicausal completely positive maps (Lemma V.5). And we characterize such generators in terms of a constructive normal form (Theorem V.6). As central insight in our proof of the latter, we use a technique based on Haar integration to

translate the semicausality property of a Lindblad generator to its completely positive part. After translating our results from semigroups of semicausal quantum operations to semigroups of quantum superchannel in Subsection V.C, we conclude our work in Section VI with an outlook on potential future research.

The idea of this project goes back to a suggestion by Michael M. Wolf and was further developed in discussions between Markus Hasenöhrl and myself. Markus Hasenöhrl is the main author of this work. In particular, he had the idea of combining Haar integration with the map Ψ_M featuring in the proof of our main result. He also developed the details of the proof strategies for Theorems IV.4, IV.10, Lemmas V.5, V.8, V.10, V.12, V.16, and Theorem V.17 in the paper, and he wrote the majority of the first draft. I was involved in all parts of the work, both regarding the mathematical content and the writing of the paper, except of the parts mentioned above.

Permission to include:

Markus Hasenöhrl and Matthias C. Caro.
Quantum and classical dynamical semigroups of superchannels and semicausal channels
Journal of Mathematical Physics 63, 072204 (2022). https://doi.org/10.1063/5.0070635

AIP Publishing LLC

Your Window to Possible

Permission to Reuse Content

REUSING AIP PUBLISHING CONTENT

Permission from AIP Publishing is required to:

- republish content (e.g., excerpts, figures, tables) if you are not the author
- modify, adapt, or redraw materials for another publication
- systematically reproduce content
- store or distribute content electronically
- copy content for promotional purposes

To request permission to reuse AIP Publishing content, use RightsLink® for the fastest response or contact AIP Publishing directly at rights@aip.org (mailto:rights@aip.org) and we will respond within one week:

For RightsLink, use Scitation to access the article you wish to license, and click on the Reprints and Permissions link under the TOOLS tab. (For assistance click the “Help” button in the top right corner of the RightsLink page.)

To send a permission request to rights@aip.org (mailto:rights@aip.org), please include the following:

- Citation information for the article containing the material you wish to reuse
- A description of the material you wish to reuse, including figure and/or table numbers
- The title, authors, name of the publisher, and expected publication date of the new work
- The format(s) the new work will appear in (e.g., print, electronic, CD-ROM)
- How the new work will be distributed and whether it will be offered for sale

Authors do **not** need permission from AIP Publishing to:

- quote from a publication (please include the material in quotation marks and provide the customary acknowledgment of the source)
- reuse any materials that are licensed under a Creative Commons CC BY license (please format your credit line: “Author names, Journal Titles, Vol.#, Article ID#, Year of Publication; licensed under a Creative Commons Attribution (CC BY) license.”)
- reuse your own AIP Publishing article in your thesis or dissertation (please format your credit line: “Reproduced from [FULL CITATION], with the permission of AIP Publishing”)
- reuse content that appears in an AIP Publishing journal for republication in another AIP Publishing journal (please format your credit line: “Reproduced from [FULL CITATION], with the permission of AIP Publishing”)
- make multiple copies of articles–although you must contact the Copyright Clearance Center (CCC) at www.copyright.com (http://www.copyright.com/) to do this

R ING CONTENT PUBLISHED BY OTHERS

content.)

Reuse of Previously Published Material Form (pdf (https://publishing.aip.org/wp-content/uploads/AIP_Permission_Form-1.pdf))

Unless the publisher requires a specific credit line, please format yours like this:

Reproduced with permission from J. Org. Chem. 63, 99 (1998). Copyright 1998, American Chemical Society.

You do not need permission to reuse material in the public domain, but you should still include an appropriate credit line which cites the original source.

Journal of Mathematical Physics

Quantum and classical dynamical semigroups of superchannels and semicausal channels

Cite as: J. Math. Phys. **63**, 072204 (2022); https://doi.org/10.1063/5.0070635
Submitted: 08 September 2021 • Accepted: 23 June 2022 • Published Online: 19 July 2022

Markus Hasenöhrl and Matthias C. Caro

Journal of Mathematical Physics ARTICLE scitation.org/journal/jmp

Quantum and classical dynamical semigroups of superchannels and semicausal channels

Cite as: J. Math. Phys. 63, 072204 (2022); doi: 10.1063/5.0070635
Submitted: 8 September 2021 • Accepted: 23 June 2022 •
Published Online: 19 July 2022

Markus Hasenöhrl[a] and Matthias C. Caro[b]

AFFILIATIONS

Department of Mathematics, Technical University of Munich, Garching, Germany and Munich Center for Quantum Science and Technology (MCQST), Munich, Germany

[a] Author to whom correspondence should be addressed: m.hasenoehrl@tum.de
[b] caro@ma.tum.de

ABSTRACT

Quantum devices are subject to natural decay. We propose to study these decay processes as the Markovian evolution of quantum channels, which leads us to dynamical semigroups of superchannels. A superchannel is a linear map that maps quantum channels to quantum channels while satisfying suitable consistency relations. If the input and output quantum channels act on the same space, then we can consider dynamical semigroups of superchannels. No useful constructive characterization of the generators of such semigroups is known. We characterize these generators in two ways: First, we give an efficiently checkable criterion for whether a given map generates a dynamical semigroup of superchannels. Second, we identify a normal form for the generators of semigroups of quantum superchannels, analogous to the Gorini-Kossakowski-Lindblad-Sudarshan form in the case of quantum channels. To derive the normal form, we exploit the relation between superchannels and semicausal completely positive maps, reducing the problem to finding a normal form for the generators of semigroups of semicausal completely positive maps. We derive a normal for these generators using a novel technique, which applies also to infinite-dimensional systems. Our work paves the way for a thorough investigation of semigroups of superchannels: Numerical studies become feasible because admissible generators can now be explicitly generated and checked. Analytic properties of the corresponding evolution equations are now accessible via our normal form.

I. INTRODUCTION AND MOTIVATION

Anybody who has ever owned an electronic device knows that these devices have a finite lifespan after which they stop working properly. At least from a consumer perspective, a long lifespan is a desirable property for such devices. Thus, it is important for an engineer to know which kind of decay processes can affect a device in order to suppress them by an appropriate design. Certainly, these considerations will also become important for the design of quantum devices. We, therefore, propose to systematically study the decay processes that quantum devices can be subject to.

In this work, we take a first step in this direction by deriving the general form of linear time-homogeneous master equations that govern how quantum channels behave when inserted into a circuit board at different points in time. This leads to the study of dynamical semigroups of superchannels. Here, superchannels are linear transformations between quantum channels.[1]

Let us consider a concrete example (see Fig. 1). Suppose we are trying to estimate the optical transmissivity of some material (M), which we assume to depend on the polarization of the incident light. A simple approach is to send photons from a light source (S) through the material and to count how many photons arrive at the detector (D). We model the material by a quantum channel T_M, acting on the states of photons described as three-level systems, with the levels corresponding to vacuum, horizontal, and vertical polarization. In an idealized world, with a perfect vacuum in the regions between the source, the material, and the detector, we can infer the transmissivity from the measurement statistics of the state $T_M(\sigma)$, where σ is the state of the photon emitted from the source. However, in a more realistic scenario, even though we might have created an (almost) perfect vacuum between the devices at construction time, some particles are leaked into that region over time.

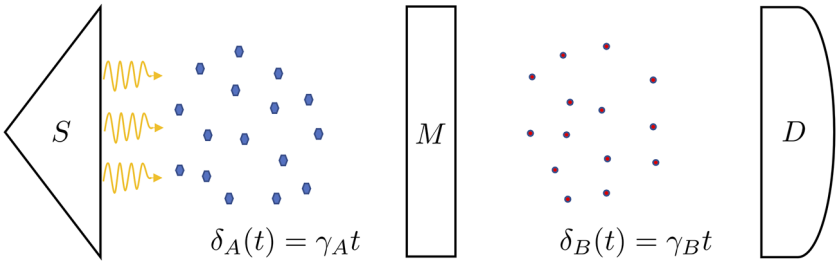


FIG. 1. Estimating the transmissivity of a material under the influence of an influx of particles into the regions between the components.

FIG. 2. If the particle density is low, then the incident photon interacts with the particles in the region sequentially and independently. The effect of a single interaction can be described by a channel ΔQ. Hence, the state after the first interaction is $\Delta Q(\sigma)$, the state after the second interaction is $\Delta Q(\Delta Q(\sigma))$, and so forth. The number of interactions is given by the product of the particle density δ and the volume V. Hence, the effect of an region with fixed volume is described by the channel $Q_\delta = (\Delta Q)^{\delta V}$. It follows that if $\delta = \delta_1 + \delta_2$, then $Q_{\delta_1+\delta_2} = (\Delta Q)^{\delta_1 V}(\Delta Q)^{\delta_2 V} = Q_{\delta_1} \circ Q_{\delta_2}$. The semigroup property for real δ can then be obtained in the continuum limit.

Then, interactions between the photons and these particles might occur, causing absorption or a change in polarization. Hence, the situation is no longer described accurately by T_M alone but also requires a description of the particle-filled regions.

To find such a description, we argue that the effect of particles in some region (here, either between S and M, or M and D) can be modeled by a quantum dynamical semigroup, parameterized by the particle density δ. If the particle density is reasonably low and Q_δ is the quantum channel describing the effect of the particles on the incident light at a given δ, then, as explained in Fig. 2, Q_δ satisfies the semigroup property $Q_{\delta_1+\delta_2} = Q_{\delta_1} \circ Q_{\delta_2}$. Furthermore, if there are no particles, then there should be no effect. Hence, $Q_0 = \mathrm{id}$. After adding continuity in the parameter δ as a further natural assumption, the family $\{Q_\delta\}_{\delta\geq 0}$ forms a quantum dynamical semigroup. That is, we can write $Q_\delta = e^{L\delta}$ for some generator L in Gorini-Kossakowski-Lindblad-Sudarshan (GKLS)-form.

If we assume in our example that particles of type A are leaked into the region between S and M at a rate γ_A and that particles of type B are leaked into the region between M and D at a rate γ_B, then the overall channel describing the transformation that emitted photons undergo at time t is given by

$$\hat{S}_t(T_M) = e^{\gamma_B L_B t} \circ T_M \circ e^{\gamma_A L_A t},$$

where L_A and L_B are the generators of the dynamical semigroups describing the effect of the particles in the respective regions.

We note that at any fixed time, $\hat{S}_t$ interpreted as a map on quantum channels is a superchannel written in "circuit"-form. This means that $\hat{S}_t$ describes a transformation of quantum channels implemented via pre- and post-processing. Furthermore, $\hat{S}_t(T_M)$ can be determined by solving the time-homogenous master equation

$$\frac{d}{dt}T(t) = \hat{L}(T(t)),$$

where $\hat{L}(T) = \gamma_A L_A \circ T + \gamma_B T \circ L_B$, with the initial condition $T(0) = T_M$. In other words, we have

$$\hat{S}_t = e^{\hat{L}t},$$

and thus, the family $\{\hat{S}_t\}_{t\geq 0}$ forms a dynamical semigroup of superchannels.

By inductive reasoning, we, thus, arrive at our central physical hypothesis: Decay-processes of quantum devices with some sort of influx are well described by dynamical semigroups of superchannels. It follows that such decay-processes can be understood by characterizing dynamical semigroups of superchannels. Such a characterization is the main goal of our work.

In particular, we aim to understand dynamical semigroups of superchannels in terms of their generators. We characterize these generators fully by providing two results: First, we give an efficiently checkable criterion for whether a given map generates a dynamical semigroup of superchannels. Second, we identify a normal form for the generators of semigroups of quantum superchannels, analogous to the GKLS form in the case of quantum channels. Interestingly, we find that the most general form of dynamical semigroups of superchannels goes beyond the simple introductory example above.

We arrive at these results through a path (see Fig. 3) that also illuminates the connection to the classical case. We start by studying dynamical semigroups of classical superchannels, which (analogously to quantum superchannels being transformations between quantum channels)

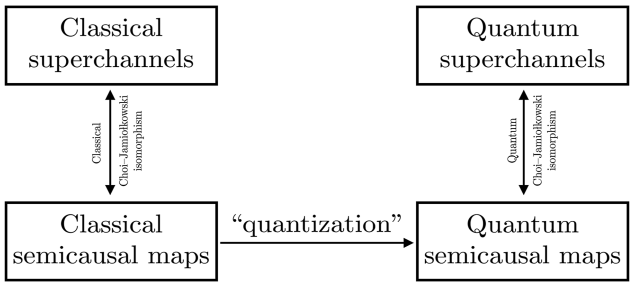


FIG. 3. Schematic of the concepts studied in this work.

are transformations between stochastic matrices. We do so by establishing a one-to-one correspondence between classical superchannels and certain classical semicausal channels, that is, stochastic matrices on a bipartite system (AB) that do not allow for communication from B to A (see Definition IV.2). We can then obtain a full characterization of the generators of semigroups of classical superchannels by characterizing generators of semigroups of classical semicausal maps first and then translating the results back to the level of superchannels. The study of (dynamical semigroups of) classical superchannels and classical semicausal channels is the content of Sec. IV.

Armed with the intuition obtained from the classical case, we then go on to study the quantum case. We start by characterizing the generators of semigroups of semicausal[2] completely positive maps (CP-maps)—our main technical result and one of independent interest. This characterization can be obtained from the classical case by a "quantization"-procedure that allows us to see exactly which features of semigroups of semicausal CP-maps are "fully quantum." Dynamical semigroups of semicausal CP-maps are discussed Sec. V B. Finally, in Sec. V C, we use the one-to-one correspondence (via the quantum Choi–Jamiołkowski isomorphism) between certain semicausal CP-maps and quantum superchannels to obtain a full characterization of the generators of semigroups of quantum superchannels. While the classical section (Sec. IV) and the quantum section (Sec. V) are heuristically related, they are logically independent and can be read independently.

This work is structured as follows: In the remainder of this section, we discuss results related to ours. Section II contains an overview over our main results. In Sec. III, we recall relevant notions from functional analysis and quantum information, as well as some notation. The (logically) independent sections (Secs. IV and V) comprise the main body of our paper, containing complete statements and proofs of our results on dynamical semigroups of superchannels and semicausal channels. We study the classical case in Sec. IV and the quantum case in Sec. V. Finally, we conclude with a summary and an outlook to future research in Sec. VI.

A. Related work

The study of quantum superchannels goes back to Ref. 1 and has since evolved to the study of higher-order quantum maps.[3–5] A peculiar feature of higher-order quantum theory is that it allows for indefinite causal order.[6,7] However, it was recently discovered that the causal order is preserved under (certain) continuous evolutions.[8,9] It, therefore, seems interesting to study continuous evolutions of higher-order quantum maps systematically. Our work can be seen as an initial step into his direction.

The study of (semi-)causal and (semi-)localizable quantum channels goes back to Ref. 2. By proving the equivalence of semicausality and semilocalizability for quantum channels, the authors of Ref. 10 resolved a conjecture raised in Ref. 2 (and attributed to DiVincenzo). Later, the authors of Ref. 11 provided an alternative proof for this equivalence and further investigated causal and local quantum operations.

II. RESULTS

We give an overview over our answers to the questions identified in Sec. I. In our first result, we identify a set of constraints that a linear map satisfies if and only if it generates a semigroup of quantum superchannels.

Result 1.1 (Lemma V.17—informal). *Checking whether a linear map* $\hat{L} : \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B)) \to \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ *generates a semigroup of quantum superchannels can be phrased as a semidefinite constraint satisfaction problem.*

Therefore, we can efficiently check whether a given linear map is a valid generator of a semigroup of quantum superchannels. We can even solve optimization problems over such generators in terms of semidefinite programs. Thereby, this first characterization of generators of semigroups of quantum superchannels facilitates working with them computationally.

As our second result, we determine a normal form for generators of semigroups of quantum superchannels. Similar to the GKLS-form, we decompose the generator into a "dissipative part" and a "Hamiltonian part," where the latter generates a semigroup of invertible superchannels such that the inverse is a superchannel as well.

Result 1.2 (Theorem V.18—informal). *A linear map* $\hat{L} : \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B)) \to \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ *generates a semigroup of quantum superchannels if and only if it can be written as* $\hat{L}(T) = \hat{D}(T) + \hat{H}(T)$*, where the "Hamiltonian part" is of the form*

$$\hat{H}(T)(\rho) = -i[H_B, T(\rho)] - iT([H_A, \rho]),$$

with local Hamiltonians H_B *and* H_A*, and where the "dissipative part" is of the form* $\hat{D}(T)(\rho) = \mathrm{tr}_E\big[\hat{D}'(T)(\rho)\big]$*, where*

$$\hat{D}'(T)(\rho) = U(T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma)A^\dagger)U^\dagger - \frac{1}{2}(T \otimes \mathrm{id}_E)(\{A^\dagger A\,,\, \rho \otimes \sigma\}) \tag{1a}$$

$$+ B(T \otimes \mathrm{id}_E)(\rho \otimes \sigma)B^\dagger - \frac{1}{2}\{B^\dagger B\,,\, (T \otimes \mathrm{id}_E)(\rho \otimes \sigma)\} \tag{1b}$$

$$+ \big[U(T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma)),\, B^\dagger\big] + \big[B\,,\, (T \otimes \mathrm{id}_E)((\rho \otimes \sigma)A^\dagger)U^\dagger\big], \tag{1c}$$

with unitary U *and arbitrary* A *and* B*.*

The "dissipative part" consists of three terms: Term (1a) itself generates a semigroup of superchannels (for $B = 0$), with the interpretation that the transformed channel $[\hat{S}_t(T)]$ arises due to the stochastic application of $T \mapsto \mathrm{tr}_E\big[U(T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma)A^\dagger)U^\dagger\big]$ at different points in time (Dyson series expansion). Term (1b) itself generates a semigroup of superchannels (for $A = 0$) of the form $\hat{S}_t(T) = e^{L_B t} \circ T$, where L_B is a generator of a quantum dynamical semigroup (and hence in GKLS-form). Term (1c) is a "superposition" term, which is harder to interpret. It will become apparent from the path taken via the "quantization" of semicausal semigroups that this term is a pure quantum feature with no classical analog. Therefore, the presence of (1c) can be regarded as one of our main findings. It is also worth noting that the normal form in Result 1.2 is more general than the form of the generator we found in our introductory example. Hence, nature allows for more general decay-processes than the simple ones with an independent influx of particles before and after the target object. We also complement this structural result by an algorithm that determines the operators U, A, B, H_A, and H_B if the conditions in Result 1.1 are met.

The proof of these results relies on the relation (via the Choi–Jamiołkowski isomorphism) between superchannels and semicausal CP-maps. Our next findings—and from a technical standpoint our main contributions—are the corresponding results for semigroups of semicausal CP-maps.

Result 2.1 (Lemma V.5—informal). *Checking whether a linear map* $L : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ *generates a semigroup of* $B \not\to A$ *semicausal CP-maps can be phrased as a semidefinite constraint satisfaction problem for its Choi-matrix.*

Based on this insight, we can efficiently check whether a given linear map is a valid generator of a semigroup of semicausal CP-maps.

Since semigroups of semicausal CP-maps are, in particular, semigroups of CP-maps, our normal form for generators giving rise to semigroups of semicausal CP-maps is a refining of the GKLS-form.

Result 2.2 (Theorem V.6—informal). *A linear map* $L : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ *generates a semigroup of* $B \not\to A$ *semicausal CP-maps (in the Heisenberg picture) if and only if it can be written as* $L(X) = \Phi(X) - K^\dagger X - XK$*, where the CP part* Φ *is of the form*

$$\Phi(X) = V^\dagger(X \otimes \mathbb{1}_E)V, \text{ with } V = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B) + (\mathbb{1}_A \otimes B),$$

with a unitary $U \in \mathcal{B}(\mathcal{H}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$ *and arbitrary* $A \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_E)$ *and* $B \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$*, and the* K *in the non-CP part is of the form*

$$K = (\mathbb{1}_A \otimes B^\dagger U)(A \otimes \mathbb{1}_B) + \frac{1}{2}\mathbb{1}_A \otimes B^\dagger B + K_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes iH_B,$$

with a self-adjoint H_B *and an arbitrary* K_A*.*

This characterization has both computational and analytical implications: On the one hand, it provides a recipe for describing semicausal GKLS generators in numerical implementations. On the other hand, the constructive characterization of semicausal GKLS generators makes a more detailed analysis of their (e.g., spectral) properties tractable. It is also worth noting that in Result 2.2, we can allow for (separable) infinite-dimensional spaces. In the finite-dimensional case, we also provide an algorithm to compute the operators U, A, B, K_A, and H_B, if the conditions of Result 2.1 are met.

Let us now turn to the corresponding results in the classical case. Here, instead of looking at (semigroups of) CP-maps and quantum channels, we look at (entry-wise) non-negative matrices and row-stochastic matrices (see Secs. III and IV for details) that we assume to act on $\mathbb{R}^{\mathbb{X}}$ for (finite) alphabets $\mathbb{X} \in \{\mathbb{A}, \mathbb{B}, \mathbb{E}\}$.

The following result is the classical analog of Result 2.2:

Result 3 (Corollary IV.8—informal). *A linear map $Q : \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}} \to \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}}$ generates a semigroup of (Heisenberg) $B \not\to A$ semicausal non-negative matrices if and only if it can be written as*

$$Q = (A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes U) - K_A \otimes \mathbb{1}_B + \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes B^{(i)},$$

with a row-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}})$, a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}}; \mathbb{R}^{\mathbb{A}})$, a diagonal matrix K_A, and maps $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ that generate semigroups of row-stochastic matrices.

We will discuss in detail how Result 2.2 arises as the "quantization" of Result 3 in the paragraph following the Proof of Lemma V.5. Here, we highlight that in both the quantum and the classical case, the generators of semicausal semigroups are constructed from two basic building blocks. In the quantum case, these are a $B \not\to A$ semicausal CP-map Φ_{sc}, with $\Phi_{sc}(X) = V_{sc}^\dagger(X \otimes \mathbb{1}_E)V_{sc}$ and $V_{sc} = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$ and a GKLS generator of the form $\mathrm{id}_A \otimes \hat{B}$. In the classical case, they are a $B \not\to A$ semicausal non-negative map $\Phi_{sc} = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$ and operators of the form $|a_i\rangle\langle a_i| \otimes B^{(i)}$, where $B^{(i)}$ generates a semigroup of row-stochastic maps. The difference between the quantum case and the classical case then lies in the way the general form is constructed from the building blocks. While we simply take convex combinations of the building-blocks in the classical case, we have to take superpositions of the building-blocks, by which we mean that we need to combine the corresponding Strinespring operators, in the quantum case.

As our last result, we present the normal form for generators of semigroups of classical superchannels.

Result 4. *A linear map $\hat{Q} : \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}}) \to \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}})$ generates a semigroup of classical superchannels if and only if it can be written as*

$$\hat{Q}(M) = U(M \otimes \mathbb{1}_E)A - \sum_{i=1}^{|\mathbb{A}|} \langle \mathbf{1}_{AE}|Aa_i\rangle M|a_i\rangle\langle a_i| + \sum_{i=1}^{|\mathbb{A}|} B^{(i)} M|a_i\rangle\langle a_i|,$$

with a column-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{B}})$, a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}})$, a diagonal matrix K_A, and a collection of generators of semigroups of column-stochastic matrices $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$.

As in the quantum case, we have two kinds of evolutions: a stochastic application of $M \mapsto U(M \otimes \mathbb{1}_E)A$ at different points in time and a conditioned post-processing evolution of the form $\sum_i e^{B^{(i)}t} M|a_i\rangle\langle a_i|$. Note that there are no "superposition" terms, such as (1c).

III. NOTATION AND PRELIMINARIES

In this section, we review basic notions from functional analysis, quantum information theory, and the theory of dynamical semigroups. We also fix our notation for these settings as well as for a classical counterpart of the quantum setting.

A. Functional analysis

Throughout this paper, $\mathcal{H}$ (with some subscript) denotes a (in general, infinite-dimensional) separable complex Hilbert space. Whenever $\mathcal{H}$ is assumed to be finite-dimensional, we explicitly state this assumption. We denote the Banach space of bounded linear operators with domain $\mathcal{H}_A$ and codomain $\mathcal{H}_B$, equipped with the operator norm, by $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$ and write $\mathcal{B}(\mathcal{H})$ for $\mathcal{B}(\mathcal{H}; \mathcal{H})$. For $X \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$, the adjoint $X^\dagger \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_A)$ of X is the unique linear operator such that $\langle \psi_B|X\psi_A\rangle = \langle X^\dagger\psi_B|\psi_A\rangle$ for all $|\psi_A\rangle \in \mathcal{H}_A$ and all $|\psi_B\rangle \in \mathcal{H}_B$. Here, and throughout this paper, we use the standard Dirac notation.

An operator $Y \in \mathcal{B}(\mathcal{H})$ is called self-adjoint if $Y^\dagger = Y$. A self-adjoint $Y \in \mathcal{B}(\mathcal{H})$ is called positive semidefinite, denoted by $Y \geq 0$, if there exists an operator $Z \in \mathcal{B}(\mathcal{H})$ such that $Y = Z^\dagger Z$. If Y is positive semidefinite, then there exists a unique positive semidefinite operator $\sqrt{Y}$ such that $Y = \sqrt{Y}\sqrt{Y}$ (Ref. 12, p. 196). The operator $\sqrt{Y}$ is called the square-root of Y. The absolute value $|Y| \in \mathcal{B}(\mathcal{H})$ of Y is defined by $|Y| = \sqrt{Y^\dagger Y}$.

We define the set of trace-class operators $\mathcal{S}_1(\mathcal{H}_A; \mathcal{H}_B) = \{\rho \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_B) \,|\, \mathrm{tr}[|\rho|] < \infty\}$, which becomes a Banach space when endowed with the norm $\|\rho\|_1 := \mathrm{tr}[|\rho|]$. We write $\mathcal{S}_1(\mathcal{H})$ for $\mathcal{S}_1(\mathcal{H}; \mathcal{H})$. The set $\mathcal{S}_1(\mathcal{H}_A; \mathcal{H}_B)$ satisfies the two-sided$*$-ideal property: If $\rho \in \mathcal{S}_1(\mathcal{H}_A; \mathcal{H}_B)$ and $Y \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$, then $\rho^\dagger \in \mathcal{S}_1(\mathcal{H}_B; \mathcal{H}_A)$, $\rho^\dagger Y \in \mathcal{S}_1(\mathcal{H}_A)$, and $Y\rho^\dagger \in \mathcal{S}_1(\mathcal{H}_B)$.

Besides the norm topology, we will use the strong operator topology and the ultraweak topology. The strong operator topology is the smallest topology on $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$ such that for all $|\psi_A\rangle \in \mathcal{H}_A$, the map $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B) \ni Y \mapsto Y|\psi_A\rangle \in \mathcal{H}_B$ is continuous, where $\mathcal{H}_B$ is equipped with the norm topology. The ultraweak topology on $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$ is the smallest topology such that the map $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B) \ni Y \mapsto \mathrm{tr}\left[\rho^\dagger Y\right] \in \mathbb{C}$ is continuous for all $\rho \in \mathcal{S}_1(\mathcal{H}_A; \mathcal{H}_B)$. Since $\mathcal{H}_A$ and $\mathcal{H}_B$ are separable, so is $\mathcal{S}_1(\mathcal{H}_B; \mathcal{H}_A)$. Hence, the sequential Banach Alaoglu theorem implies that every bounded sequence in $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$ has an ultraweakly convergent subsequence. Here, we view $\mathcal{B}(\mathcal{H}_A; \mathcal{H}_B)$ as the continuous dual of $\mathcal{S}_1(\mathcal{H}_B; \mathcal{H}_A)$. The aforementioned results can be found in many books, e.g., Ref. 12 (ch. VI.6), however, usually only for the case $\mathcal{H}_A = \mathcal{H}_B$.

The general results stated above can be obtained from this case by considering $\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B)$ and $\mathcal{S}_1(\mathcal{H}_A;\mathcal{H}_B)$ as subspaces of $\mathcal{B}(\mathcal{H}_A\oplus\mathcal{H}_B)$ and $\mathcal{S}_1(\mathcal{H}_A\oplus\mathcal{H}_B)$, respectively.

An operator $V\in\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B)$ is called an isometry if $\|V|\psi_A\rangle\| = \||\psi_A\rangle\|$ for all $|\psi_A\rangle\in\mathcal{H}_A$. The (possibly empty) set of unitaries, the surjective isometries, is denoted by $\mathcal{U}(\mathcal{H}_A;\mathcal{H}_B)$, and we write $\mathcal{U}(\mathcal{H})$ for $\mathcal{U}(\mathcal{H};\mathcal{H})$. As a special notation, if $\mathcal{H}'_A$ and $\mathcal{H}'_B$ are closed linear subspaces of $\mathcal{H}_A$ and $\mathcal{H}_B$, with (canonical) isometric embeddings $\mathbb{1}_{A'\to A}\in\mathcal{B}(\mathcal{H}'_A;\mathcal{H}_A)$ and $\mathbb{1}_{B'\to B}\in\mathcal{B}(\mathcal{H}'_B;\mathcal{H}_B)$, respectively, then we will write $\mathcal{U}_P(\mathcal{H}'_A;\mathcal{H}'_B) = \{\mathbb{1}_{B'\to B}U\mathbb{1}^\dagger_{A'\to A}\in\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B)\,|\,U\in\mathcal{U}(\mathcal{H}'_A;\mathcal{H}'_B)\}$ and $\mathcal{U}_P(\mathcal{H})$ for $\mathcal{U}_P(\mathcal{H};\mathcal{H})$. That is, this is the set of partial isometries.

B. Flip operator, partial trace, complete positivity, and duality

The flip operator $\mathbb{F}_{A;B}\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_A)$ is the unique operator satisfying $\mathbb{F}_{A;B}(|\psi_A\rangle\otimes|\psi_B\rangle) = |\psi_B\rangle\otimes|\psi_A\rangle$ for all $|\psi_A\rangle\in\mathcal{H}_A$ and all $|\psi_B\rangle\in\mathcal{H}_B$.

The partial trace with respect to the space $\mathcal{H}_A$ is the unique linear map $\mathrm{tr}_A:\mathcal{S}_1(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)\to\mathcal{S}_1(\mathcal{H}_B;\mathcal{H}_C)$ that satisfies $\mathrm{tr}[X\mathrm{tr}_A[\rho]] = \mathrm{tr}[(\mathbb{1}_A\otimes X)\rho]$ for all $\rho\in\mathcal{S}_1(\mathcal{H}_A\otimes\mathcal{H}_B)$ and all $X\in\mathcal{B}(\mathcal{H}_C;\mathcal{H}_B)$. If the spaces involved have subscripts, the partial trace will always be denoted with the corresponding subscript. The partial trace with respect to $\rho\in\mathcal{S}_1(\mathcal{H}_A)$ is the unique linear map $\mathrm{tr}_\rho:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)\to\mathcal{B}(\mathcal{H}_B;\mathcal{H}_C)$ that satisfies $\mathrm{tr}[\sigma\mathrm{tr}_\rho[X]] = \mathrm{tr}[(\rho\otimes\sigma)X]$ for all $\sigma\in\mathcal{S}_1(\mathcal{H}_C;\mathcal{H}_B)$ and all $X\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)$. Proofs of existence and uniqueness can be found in Ref. 13 (Theorem 2.28 and Theorem 2.30), where we used again the observation that the results above follow from the usual ones for $\mathcal{H}_B=\mathcal{H}_C$, by looking at the operators on $\mathcal{H}_A\otimes(\mathcal{H}_B\oplus\mathcal{H}_C)$.

Let $T\in\mathcal{B}(\mathcal{B}(\mathcal{H}_B);\mathcal{B}(\mathcal{H}_A))$. The map T is called positive if $T(X_B)$ is positive semidefinite whenever $X_B\in\mathcal{B}(\mathcal{H}_B)$ is positive semidefinite. For $n\in\mathbb{N}_0$, the map $T_n:\mathcal{B}(\mathbb{C}^n\otimes\mathcal{H}_B)\to\mathcal{B}(\mathbb{C}^n\otimes\mathcal{H}_A)$ is uniquely defined by the requirement that $T_n(X_n\otimes X_B) = X_n\otimes T(X_B)$ for all $X_n\in\mathcal{B}(\mathbb{C}^n)$ and all $X_B\in\mathcal{B}(\mathcal{H}_B)$. The map T is completely positive (CP) if the map T_n is positive for all $n\in\mathbb{N}_0$. A CP-map T is called normal if T is continuous when $\mathcal{B}(\mathcal{H}_A)$ and $\mathcal{B}(\mathcal{H}_B)$ are both equipped with the ultraweak topology. We denote the set of normal CP-maps by $\mathrm{CP}_\sigma(\mathcal{H}_B;\mathcal{H}_A)$ and write $\mathrm{CP}_\sigma(\mathcal{H})$ for $\mathrm{CP}_\sigma(\mathcal{H};\mathcal{H})$. By the Stinespring dilation theorem (in its form for normal CP-maps), T is a normal CP-map if and only if there exist a (separable) Hilbert space $\mathcal{H}_E$ and an operator $V\in\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B\otimes\mathcal{H}_E)$ such that for all $X_B\in\mathcal{B}(\mathcal{H}_B)$, we have $T(X_B) = V^\dagger(X_B\otimes\mathbb{1}_E)V$. Furthermore, the Stinespring dilation can be chosen to be minimal, that is, the pair $(V,\mathcal{H}_E)$ can be chosen such that $\mathrm{span}\{(X_B\otimes\mathbb{1}_E)V|\psi_A\rangle\,|\,X_B\in\mathcal{B}(\mathcal{H}_B),|\psi_A\rangle\in\mathcal{H}_A\}$ is norm-dense in $\mathcal{H}_B\otimes\mathcal{H}_E$. Furthermore, if $(V',\mathcal{H}'_E)$ is another Stinespring dilation, then there exists an isometry $U\in\mathcal{B}(\mathcal{H}_E;\mathcal{H}'_E)$ such that $V' = (\mathbb{1}_B\otimes U)V$. Another equivalent characterization is the so-called Kraus form: T is a normal CP-map if and only if there exists a countable set of operators $\{L_i\}_i\subset\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B)$, the Kraus operators, such that for all $X_B\in\mathcal{B}(\mathcal{H}_B)$, we have $T(X_B) = \sum_i L_i^\dagger X_B L_i$, where the series converges in the strong operator topology. One can obtain Kraus operators from a Stinespring dilation $(V,\mathcal{H}_E)$ by choosing an orthonormal basis $\{|e_i\rangle\}_i$ of $\mathcal{H}_E$ and defining $L_i = (\mathbb{1}_B\otimes\langle e_i|)V$. A map T is unital if $T(\mathbb{1}_B) = \mathbb{1}_A$, and a unital normal CP-map is called a Heisenberg (quantum) channel.

Let $S\in\mathcal{B}(\mathcal{S}_1(\mathcal{H}_A);\mathcal{S}_1(\mathcal{H}_B))$. The dual map $S^*\in\mathcal{B}(\mathcal{B}(\mathcal{H}_B);\mathcal{B}(\mathcal{H}_A))$ is the unique linear map that satisfies $\mathrm{tr}[X_B^\dagger S(\rho)] = \mathrm{tr}[\big(S^*(X_B)\big)^\dagger\rho]$ for all $X_B\in\mathcal{B}(\mathcal{H}_B)$ and all $\rho\in\mathcal{S}_1(\mathcal{H}_A)$. We call S the Schrödinger picture map and S^* the Heisenberg picture map. The map S is called completely positive if S^* is completely positive in the sense defined above. In that case, S^* is automatically normal. In fact, T is a normal CP-map if and only if there exists $S\in\mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ such that $S^* = T$. It follows that S is completely positive if and only if there exist a separable Hilbert space $\mathcal{H}_E$ and an operator $V\in\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B\otimes\mathcal{H}_E)$ such that $S(\rho) = \mathrm{tr}_E\big[V\rho V^\dagger\big]$ for all $\rho\in\mathcal{S}_1(\mathcal{H}_A)$. Furthermore, S is completely positive if and only if there exist a countable set of operators $\{L_i\}_i\subset\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B)$ such that $S(\rho) = \sum_i L_i\rho L_i^\dagger$ and the series converges in trace-norm. A map S is trace-preserving if $\mathrm{tr}[S(\rho_A)] = \mathrm{tr}[\rho_A]$ for all $\rho_A\in\mathcal{S}_1(\mathcal{H}_A)$. A trace-preserving CP-map is called a (quantum) channel. The facts in this section are contained or follow directly from the results in Refs. 14 and 15.

C. Choi–Jamiołkowski isomorphism, partial transposition

In this section, let $\mathcal{H}_A$, $\mathcal{H}_B$, and $\mathcal{H}_C$ be finite-dimensional Hilbert spaces with fixed orthonormal bases $\{|a_i\rangle\}_i$, $\{|b_j\rangle\}_j$, and $\{|c_k\rangle\}_k$, respectively. The transpose (with respect to $\{|a_i\rangle\}_i$ and $\{|b_j\rangle\}_j$) of an operator $X\in\mathcal{B}(\mathcal{H}_A;\mathcal{H}_B)$ is the unique linear operator $X^T\in\mathcal{B}(\mathcal{H}_B;\mathcal{H}_A)$ such that $\langle b_j|Xa_i\rangle = \langle a_i|X^Tb_j\rangle$ for all elements of the orthonormal bases. The partial transposition (with respect to $\{|a_i\rangle\}_i$) of an operator $X\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)$ is the unique linear operator $X^{T_A}\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)$ such that $(\langle a_i|\otimes\mathbb{1}_C)X(|a_j\rangle\otimes\mathbb{1}_B) = (\langle a_j|\otimes\mathbb{1}_C)X^{T_A}(|a_i\rangle\otimes\mathbb{1}_B)$ for all elements of the orthonormal basis.

The (quantum) Choi–Jamiołkowski isomorphism,[16,17] defined with respect to an orthonormal basis $\{|a_i\rangle\}_i$ of $\mathcal{H}_A$, is the bijective linear map $\mathfrak{C}_{A;B}:\mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$, $\mathfrak{C}_{A;B}(T) = (\mathrm{id}_A\otimes T)(|\Omega\rangle\langle\Omega|)$, and its inverse is given by $\mathfrak{C}^{-1}_{A;B}(\tau)(\rho) = \mathrm{tr}_A\big[(\rho^T\otimes\mathbb{1})\tau\big]$, where $|\Omega\rangle := \sum_i|a_i\rangle\otimes|a_i\rangle$. A map $S\in\mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ is completely positive if and only if $\mathfrak{C}_{A;B}(S)\geq 0$; S is trace-preserving if and only if $\mathrm{tr}_B[\mathfrak{C}_{A;B}(S)] = \mathbb{1}_A$, and we have the identity $\mathrm{tr}_A[\mathfrak{C}_{A;B}(S)] = S(\mathbb{1}_A)$. We will occasionally call elements of the image of $\mathfrak{C}_{A;B}$ Choi matrices.

D. Non-negative matrices and duality

As we provide characterizations for both the quantum and the classical case, we now also introduce the notation and definitions required for the latter. With a classical system A, we associate a finite alphabet $\mathbb{A} = \{a_1,a_2,\ldots,a_{|\mathbb{A}|}\}$ and a "state-space" $\mathbb{R}^{\mathbb{A}}$, with the orthonormal basis

$\{|a_i\rangle\}_{i=1}^{|\mathbb{A}|}$. We define by $|\mathbf{1}_A\rangle := \sum_i |a_i\rangle$ the all-one-vector. A vector $|x\rangle \in \mathbb{R}^{\mathbb{A}}$ is called non-negative if $\langle a|x\rangle \geq 0$ for all $a \in \mathbb{A}$. A linear operator $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}})$ is called non-negative if $M|x\rangle$ is non-negative whenever $|x\rangle$ is non-negative (equivalently, all matrix elements are non-negative). A non-negative M is called column-stochastic if $\langle \mathbf{1}_B|M = \langle \mathbf{1}_A|$, column-sub-stochastic if there exists a non-negative P such that $M + P$ is column-stochastic, row-stochastic if $M|\mathbf{1}_A\rangle = |\mathbf{1}_B\rangle$, and row-sub-stochastic if there exists a non-negative P such that $M + P$ is row-stochastic. Given $|x\rangle$ or $\langle x|$, we denote by $\mathrm{diag}(|x\rangle) = \mathrm{diag}(\langle x|)$ the diagonal matrix with the components of x on the diagonal. Finally, we will use the "classical Choi–Jamiołkowski isomorphism" (also known as vectorization), which is a convenient notation to make the connection to the quantum case more transparent. The classical Choi–Jamiołkowski isomorphism, defined with respect to $\{|a_i\rangle\}_i$, is the linear map $\mathfrak{C}^C_{A;B} : \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}}) \to \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ defined by $\mathfrak{C}^C_{A;B}(M) = (\mathbb{1}_A \otimes M)|\Omega\rangle$, where $|\Omega\rangle := \sum_i |a_i\rangle \otimes |a_i\rangle$. The inverse $(\mathfrak{C}^C_{A;B})^{-1}$ is then given by $(\mathfrak{C}^C_{A;B})^{-1}(|x\rangle) = (\langle\Omega| \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes |x\rangle)$ We will sometimes refer to elements of the range of $\mathfrak{C}^C_{A;B}$ as Choi vectors.

E. Dynamical semigroups

Let $\mathcal{X}$ be a Banach space. A family of operators $\{T_t\}_{t\geq 0}$, with $T_t \in \mathcal{B}(\mathcal{X})$ for all $t \geq 0$, is called a norm-continuous one-parameter semigroup on $\mathcal{X}$ or, short, dynamical semigroup if $T_0 = \mathbb{1}$, $T_{s+t} = T_s T_t$ for all $t, s \geq 0$ and the map $\mathbb{R}_{\geq 0} \ni t \mapsto T_t$ is norm-continuous. Norm-continuous dynamical semigroups are automatically differentiable and have bounded generators, that is, there exists $L \in \mathcal{B}(\mathcal{X})$ such that $T_t = e^{tL}$ for all $t \geq 0$ and $L = \frac{d}{dt}\big|_{t=0_+} T_t$ (Ref. 18, Theorem I.3.7).

Lindblad[19] proved that $T_t \in \mathrm{CP}_\sigma(\mathcal{H})$ for all $t \geq 0$ if and only if there exist $\Phi \in \mathrm{CP}_\sigma(\mathcal{H})$ and $K \in \mathcal{B}(\mathcal{H})$ such that $T_t = e^{tL}$, with $L(X) = \Phi(X) - K^\dagger X - XK$. In this case, we refer to $\{T_t\}_{t\geq 0}$ as a CP semigroup. We call the corresponding form of the generator L the GKLS form[19,20] and Φ its CP part. If $\mathcal{H}$ is finite-dimensional, then $T_t = e^{tL} \in \mathrm{CP}_\sigma(\mathcal{H})$ for all $t \geq 0$ if and only if the operator $\mathfrak{L} := \mathfrak{C}_{A;B} = (\mathrm{id} \otimes L)(|\Omega\rangle\langle\Omega|)$ is self-adjoint and $P^\perp \mathfrak{L} P^\perp \geq 0$, where $|\Omega\rangle = \sum_i |a_i\rangle \otimes |a_i\rangle$ for some orthonormal basis $\{a_i\}$ of $\mathcal{H}$ and $P^\perp \in \mathcal{B}(\mathcal{H} \otimes \mathcal{H})$ is the orthogonal projection onto the orthogonal complement of $\{|\Omega\rangle\}$.[21,22] The corresponding classical result is as follows: $\{T_t\}_{t\geq 0} \subseteq \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ is a dynamical semigroup of non-negative linear maps if and only if there exist a non-negative linear map $\Phi \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ and a diagonal map $K \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ (with respect to the basis orthogonal basis $\{|a_i\rangle\}_i$) such that the generator L has the form $\Phi - K$.[23]

IV. THE CLASSICAL CASE

Before studying the quantum scenario, we consider the classical version of our main question. That is, we study continuous semigroups of classical superchannels and their generators. On the one hand, this allows us to develop an intuition that we can build upon for the quantum case. On the other hand, a comparison between the classical and the quantum case elucidates which features of the latter are actually quantum. For the purpose of this section, $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{E}$ denote finite alphabets as in Subsection III D.

A classical superchannel is a map that maps classical channels, i.e., stochastic matrices, to classical channels while preserving the probabilistic structure of the classical theory. To achieve the latter requirement, we require that a classical superchannel is a linear map and that probabilistic transformations, i.e., sub-stochastic matrices, are mapped to probabilistic transformations. Expressed more formally, we have the following definition:

Definition IV.1 (classical superchannels). A linear map $\hat{S} : \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}}) \to \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}})$ is called a classical superchannel if $\hat{S}(M) \in \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}})$ is column sub-stochastic whenever $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}})$ is column sub-stochastic and $\hat{S}(M) \in \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}})$ is column stochastic whenever $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}};\mathbb{R}^{\mathbb{B}})$ is column stochastic.

A related concept is that of a classical semicausal channel, which is a stochastic matrix on a bipartite space $\mathbb{A} \times \mathbb{B}$ such that no communication from B to A is allowed. We formalize this as follows:

Definition IV.2 (classical semicausality). An operator $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is called column $B \not\to A$ semicausal if there exists $M^A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ such that $(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)M = M^A(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)$.
Similarly, $N \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is called row $B \not\to A$ semicausal if there exists $N^A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ such that $N(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) = N^A \otimes |\mathbf{1}_B\rangle$.

Clearly, M is column $B \not\to A$ semicausal if and only if M^T is row $B \not\to A$ semicausal. To emphasize the analogy to the quantum case, we will often refer to a column $B \not\to A$ semicausal map as a Schrödinger $B \not\to A$ semicausal map and to a row $B \not\to A$ semicausal map as a Heisenberg $B \not\to A$ semicausal map. In both cases, the maps M^A and N^A will be called the reduced maps.

The structure of this section is as follows: We start by establishing the connection between classical superchannels and classical non-negative semicausal maps, followed by a characterization of classical non-negative semicausal maps as a composition of known objects; such a characterization is known in the quantum case as the equivalence between semicausality and semilocalizability. We then turn to the study of the generators of semigroups of semicausal and non-negative maps and finally use the correspondence between superchannels and semicausal channels to obtain the corresponding results for the generators of semigroups of superchannels.

A. Correspondence between classical superchannels and semicausal non-negative linear maps

We first show, with a proof inspired by the one given in Ref. 1 for the analogous correspondence in the quantum case, that we can understand classical superchannels in terms of classical semicausal channels. To concisely state this correspondence, we use the classical

version of the Choi–Jamiołkowski isomorphism. Let us mention here once again that we assume all alphabets $(\mathbb{A}, \mathbb{B}, \dots)$ to be finite for our treatment of the classical case.

Theorem IV.3. *Let $\hat{S} : \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}}) \to \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}})$ be a linear map and define $S \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ via $S = \mathfrak{C}^C_{A;B} \circ \hat{S} \circ (\mathfrak{C}^C_{A;B})^{-1}$. Then, $\hat{S}$ is a classical superchannel if and only if S is non-negative and (Schrödinger $B \not\to A$) semicausal such that the reduced map S^A satisfies $S^A|\mathbf{1}_A\rangle = |\mathbf{1}_A\rangle$. In this case, S^A is automatically non-negative.*

Proof. We first show the "if"-direction, i.e., that if S is non-negative and (Schrödinger $B \not\to A$) semicausal, then $\hat{S} = (\mathfrak{C}^C_{A;B})^{-1} \circ S \circ \mathfrak{C}^C_{A;B}$ is a superchannel. Suppose M is a non-negative matrix. Then, $\hat{S}(M)$ is non-negative, since $\mathfrak{C}^C_{A;B}$ maps non-negative matrices to non-negative vectors, S maps non-negative vectors to non-negative vectors, and $(\mathfrak{C}^C_{A;B})^{-1}$ maps non-negative vectors to non-negative matrices.

Furthermore, if M is column stochastic, then

$$\begin{aligned}
\langle \mathbf{1}_B|\hat{S}(M) &= \langle \mathbf{1}_B|\big(\mathfrak{C}^C_{A;B}\big)^{-1} \circ S \circ \mathfrak{C}^C_{A;B}(M) \\
&= \big(\langle \Omega| \otimes \langle \mathbf{1}_B|\big)\big(\mathbb{1}_A \otimes S\big(\mathfrak{C}^C_{A;B}(M)\big)\big) \\
&= \langle \Omega|\big(\mathbb{1}_A \otimes S^A\big(\big(\mathbb{1}_A \otimes \langle \mathbf{1}_B|\big)\mathfrak{C}^C_{A;B}(M)\big)\big) \\
&= \langle \Omega|\big(\mathbb{1}_A \otimes S^A\big(\big(\mathbb{1}_A \otimes \big(\langle \mathbf{1}_B|M\big)\big)|\Omega\rangle\big)\big) \\
&= \langle \Omega|\big(\mathbb{1}_A \otimes S^A|\mathbf{1}_A\rangle\big) \\
&= \langle \Omega|\big(\mathbb{1}_A \otimes |\mathbf{1}_A\rangle\big) \\
&= \langle \mathbf{1}_A|,
\end{aligned}$$

so $\hat{S}(M)$ is stochastic. In the preceding calculation, we used that S is semicausal in the third line, that M is stochastic in the fifth line, and that $S^A|\mathbf{1}_A\rangle = |\mathbf{1}_A\rangle$ in the sixth line.

Now suppose that M is sub-stochastic such that $M + Q$ is stochastic, with Q being non-negative. Then, $\hat{S}(M + Q) = \hat{S}(M) + \hat{S}(Q)$ is stochastic, and since $\hat{S}(Q)$ is non-negative, $\hat{S}(M)$ is sub-stochastic. This proves that $\hat{S}$ is a superchannel. The claim about the non-negativity of S^A now follows directly from the semicausality condition.

For the converse, suppose $\hat{S}$ is a superchannel. Since for all $a \in \mathbb{A}$ and all $b \in \mathbb{B}$, the matrix $|b\rangle\langle a|$ is sub-stochastic, it follows by linearity of $\hat{S}$ that $\hat{S}(M)$ is non-negative whenever M is non-negative. Thus, since $(\mathfrak{C}^C_{A;B})^{-1}$ maps non-negative vectors to non-negative matrices, $\hat{S}$ maps non-negative matrices to non-negative matrices, and $\mathfrak{C}^C_{A;B}$ maps non-negative matrices to non-negative vectors, it follows that S is non-negative.

Next, we want to show that S is Schrödinger $B \not\to A$ semicausal. Since $\hat{S}$ is a superchannel, S maps Choi vectors of stochastic matrices to Choi vectors of stochastic matrices, that is, $(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S|x\rangle = |\mathbf{1}_A\rangle$ for all non-negative vectors $|x\rangle \in \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}}$ that satisfy $(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x\rangle = |\mathbf{1}_A\rangle$. As a tool, we define the set of scaled differences of Choi vectors of stochastic matrices by

$$C_0 := \big\{\lambda(|p\rangle - |n\rangle) \,|\, \lambda \in \mathbb{R};\ |p\rangle, |n\rangle \in \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}} \text{ non} - \text{negative, with } (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|p\rangle = (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|n\rangle = |\mathbf{1}_A\rangle\big\}. \tag{2}$$

We claim that

$$C_0 = C_0' := \big\{|x'\rangle \in \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}} \,|\, (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x'\rangle = 0\big\}.$$

To see this, first note that $C_0 \subseteq C_0'$ follows directly from the definition. For the other inclusion, $C_0 \supseteq C_0'$, we decompose $|x'\rangle \in C_0'$ as $|x'\rangle = |p'\rangle - |n'\rangle$ for two non-negative vectors $|p'\rangle, |n'\rangle \in \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}}$. It follows that $(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|p'\rangle = (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|n'\rangle$. Furthermore, for $\varepsilon > 0$ small enough, we have that $|y'\rangle := |\mathbf{1}_A\rangle - \varepsilon(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|p'\rangle$ is non-negative. However, for any non-negative unit $|v\rangle \in \mathbb{R}^{\mathbb{B}}$, with $\langle \mathbf{1}_B|v\rangle = 1$, the vectors $|p\rangle := \varepsilon|p'\rangle + |y'\rangle \otimes |v\rangle$ and $|n\rangle := \varepsilon|n'\rangle + |y'\rangle \otimes |v\rangle$ are Choi vectors of stochastic matrices. Hence, $|x'\rangle = \frac{1}{\varepsilon}(|p\rangle - |n\rangle) \in C_0$.

We define $P^\perp \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ by $P^\perp|x\rangle = \frac{1}{|\mathbb{B}|}\big[(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x\rangle\big] \otimes |\mathbf{1}_B\rangle$ and $P := \mathbb{1}_{AB} - P^\perp$. Then, since $(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)P|x\rangle = (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x\rangle - (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x\rangle = 0$, we have that $P|x\rangle \in C_0$ for all $|x\rangle \in \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}}$. We define $S^A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ by $S^A|x_A\rangle = \frac{1}{|\mathbb{B}|}(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)P^\perp S(|x_A\rangle \otimes |\mathbf{1}_B\rangle)$ $= \frac{1}{|\mathbb{B}|}(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S(|x_A\rangle \otimes |\mathbf{1}_B\rangle)$ and calculate

$$\begin{aligned}
(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S|x\rangle &= (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S(P|x\rangle) + (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S(P^\perp|x\rangle) \\
&= (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S(P^\perp|x\rangle) \\
&= (\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S\left(\frac{1}{|\mathbb{B}|}\big[(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x\rangle\big] \otimes |\mathbf{1}_B\rangle\right) \\
&= S^A((\mathbb{1}_A \otimes \langle \mathbf{1}_B|)|x\rangle),
\end{aligned}$$

where we used in the second line that C_0 is invariant under S, a fact that follows directly from (2). This calculation exactly shows that S is Schödinger $A \not\to B$ semicausal.

It remains to show that $S^A|\mathbf{1}_A\rangle = |\mathbf{1}_A\rangle$. This follows easily, since

$$\begin{aligned} S^A|\mathbf{1}_A\rangle &= \frac{1}{|\mathbb{B}|}(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)S(|\mathbf{1}_A\rangle \otimes |\mathbf{1}_B\rangle) \\ &= \frac{1}{|\mathbb{B}|}(\mathbb{1}_A \otimes \langle \mathbf{1}_B|)\mathfrak{C}^C_{A;B} \circ \hat{S} \circ (\mathfrak{C}^C_{A;B})^{-1}(|\mathbf{1}_A\rangle \otimes |\mathbf{1}_B\rangle) \\ &= \mathbb{1}_A \otimes \left[\langle \mathbf{1}_B|\hat{S}\left(\frac{1}{|\mathbb{B}|}|\mathbf{1}_B\rangle\langle \mathbf{1}_A|\right)\right]|\Omega\rangle \\ &= (\mathbb{1}_A \otimes \langle \mathbf{1}_A|)|\Omega\rangle \\ &= |\mathbf{1}_A\rangle, \end{aligned}$$

where we used that $\frac{1}{|\mathbb{B}|}|\mathbf{1}_B\rangle\langle \mathbf{1}_A|$ is stochastic and that thus $\hat{S}(\frac{1}{|\mathbb{B}|}|\mathbf{1}_B\rangle\langle \mathbf{1}_A|)$ is stochastic. □

In summary, Theorem IV.3 tells us that, via the classical Choi–Jamiołkowski isomorphism, we can view classical superchannels equivalently also as suitably normalized semicausal non-negative maps.

B. Relation between classical semicausality and semilocalizability

The goal of this section is to get a better understanding of the structure of semicausal maps. For non-negative semicausal maps, we have the following structure theorem:

Theorem IV.4. *A non-negative map $N \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is row $B \not\to A$ semicausal if and only if there exist a (finite) alphabet $\mathbb{E}$, a (non-negative) row-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}})$, and a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}}; \mathbb{R}^{\mathbb{A}})$ such that*

$$N = (A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes U). \tag{3}$$

In that case, we can choose $|\mathbb{E}| = |\mathbb{A}|^2$.

Borrowing the terminology from the quantum case,[2,10] the preceding theorem tells us that non-negative semicausal maps are semilocalizable. We formally define the latter notion for the classical case as follows:

Definition IV.5. A non-negative map $N \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is called Heisenberg $B \not\to A$ semilocalizable if it can be written in the form of Eq. (3).

Similarly, a non-negative map $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is called Schrödinger $B \not\to A$ semilocalizable if it can be written as $M = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$ for a (non-negative) column-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{B}})$ and a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}})$.

The requirement that U is stochastic and A is non-negative in the decomposition above is essential. In fact, if one drops these requirements, then a decomposition $M = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$ can be found for any matrix $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$.

Due to Theorem IV.4, a non-negative Schrödinger $B \not\to A$ semicausal and column-stochastic map M admits an operational interpretation. First, note that if M is not only semicausal but also stochastic, then also the matrix A in Eq. (3) is stochastic. Thus, the interpretation of the decomposition is as follows: First, Alice applies some probabilistic operation (A) to the composite system $\mathbb{A} \times \mathbb{E}$. Then, she transmits the E-part to Bob, who now applies a stochastic operation (U) to his part of the system.

Given this interpretation, the idea behind the construction in the Proof of Theorem IV.4 is that Alice first looks the input of system A and generates the output of system A according to the distribution given by the matrix N^A. Then, she copies the input as well as her generated output and sends this information to Bob, who is then able to complete the operation by generating an output conditional on his input and the information he got from Alice. Given that this construction requires copying, it might be considered surprising that a quantum analog is true nevertheless.[10]

Proof (Theorem IV.4). If N is Schrödinger $B \not\to A$ semilocalizable, then

$$N(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) = (A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes U|\mathbf{1}_B\rangle) = (A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes |\mathbf{1}_{EB}\rangle) = (A(\mathbb{1}_A \otimes |\mathbf{1}_E\rangle)) \otimes |\mathbf{1}_B\rangle.$$

Hence, N is row $B \not\to A$ semicausal.

Conversely, if N is row $B \not\to A$ semicausal, we choose $\mathbb{E} := \mathbb{A} \times \mathbb{A}$ and define

$$A := \sum_{i,j,k} \langle a_j | N^A a_k \rangle \; |a_j\rangle\langle a_k| \otimes \langle a_k \otimes a_j|,$$
$$U := \sum_{\substack{m,n,r,s \\ \langle a_n|N^A a_m\rangle \neq 0}} \frac{\langle a_n \otimes b_r | N\, a_m \otimes b_s\rangle}{\langle a_n | N^A a_m\rangle} |a_m \otimes a_n \otimes b_r\rangle\langle b_s| + \left[\sum_{\substack{m,n \\ \langle a_n|N^A a_m\rangle = 0}} |a_m \otimes a_n\rangle \right] \otimes \mathbb{1}_B. \tag{4}$$

To show that $N = (A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes U)$, we calculate

$$\begin{aligned}
(A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes U) &= \sum_{\substack{i,j,k \\ m,n,r,s \\ \langle a_n|N^A a_m\rangle \neq 0}} \frac{\langle a_j|N^A a_k\rangle \langle a_n \otimes b_r | N\, a_m \otimes b_s\rangle}{\langle a_n|N^A a_m\rangle} [(|a_j\rangle\langle a_k| \otimes \langle a_k \otimes a_j| \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes |a_m \otimes a_n \otimes b_r\rangle\langle b_s|)] \\
&\quad + \sum_{\substack{i,j,k \\ m,n \\ \langle a_n|N^A a_m\rangle = 0}} \langle a_j|N^A a_k\rangle (|a_j\rangle\langle a_k| \otimes \langle a_k \otimes a_j| \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes |a_m \otimes a_n\rangle \otimes \mathbb{1}_B) \\
&= \sum_{\substack{i,j,k,r,s \\ \langle a_j|N^A a_k\rangle \neq 0}} \frac{\langle a_j|N^A a_k\rangle \langle a_j \otimes b_r | N\, a_k \otimes b_s\rangle}{\langle a_j|N^A a_k\rangle} \; |a_j\rangle\langle a_k| \otimes |b_r\rangle\langle b_s| \\
&\quad + \sum_{\substack{i,j,k \\ \langle a_j|N^A a_k\rangle = 0}} \langle a_j|N^A a_k\rangle |a_j\rangle\langle a_k| \otimes \mathbb{1}_B \\
&= N.
\end{aligned}$$

For the last step, observe that the second sum vanishes and that one can drop the constraint that $\langle a_j|N^A a_k\rangle \neq 0$ in the first sum (after cancellation) because $\langle a_j \otimes b_r|N\, a_k \otimes b_s\rangle = 0$ if $\langle a_j|N^A a_k\rangle = 0$. To see this last claim, note that, since N is non-negative and semicausal, we have

$$0 \leq \langle a_j \otimes b_r|N\, a_k \otimes b_s\rangle \leq \langle a_j \otimes b_r|N\, a_k \otimes \mathbf{1}_B\rangle = \langle a_j|N^A a_k\rangle\langle b_r|\mathbf{1}_B\rangle = 0.$$

It is clear that A and U are non-negative since N and, thus, also N^A are non-negative by assumption. It remains to show that U is row-stochastic. We have

$$\begin{aligned}
U|\mathbf{1}_B\rangle &= \sum_{\substack{m,n,r,s \\ \langle a_n|N^A a_m\rangle \neq 0}} \frac{\langle a_n \otimes b_r | N\, a_m \otimes b_s\rangle}{\langle a_n|N^A a_m\rangle} |a_m \otimes a_n \otimes b_r\rangle + \sum_{\substack{m,n,s \\ \langle a_n|N^A a_m\rangle = 0}} |a_m \otimes a_n \otimes b_s\rangle \\
&= \sum_{\substack{m,n,r \\ \langle a_n|N^A a_m\rangle \neq 0}} \frac{\langle a_n \otimes b_r | N\, a_m \otimes \mathbf{1}_B\rangle}{\langle a_n|N^A a_m\rangle} |a_m \otimes a_n \otimes b_r\rangle + \sum_{\substack{m,n,s \\ \langle a_n|N^A a_m\rangle = 0}} |a_m \otimes a_n \otimes b_s\rangle \\
&= \sum_{\substack{m,n,r \\ \langle a_n|N^A a_m\rangle \neq 0}} |a_m \otimes a_n \otimes b_r\rangle + \sum_{\substack{m,n,s \\ \langle a_n|N^A a_m\rangle = 0}} |a_m \otimes a_n \otimes b_s\rangle \\
&= |\mathbf{1}_{EB}\rangle,
\end{aligned}$$

where we used the condition that N is semicausal to obtain the third line. This finishes the proof. □

Remark IV.6. Theorem IV.4 can be extended to weak $-^*$ continuous non-negative maps on the Banach space of bounded real sequences, but this requires extra care and does not yield additional insight beyond the previous proof.

C. Generators of semigroups of classical semicausal non-negative maps

The main goal of this section is to establish a structure theorem for the generators of semigroups of non-negative semicausal maps. First, recall that a (norm)-continuous semigroup $\{N_t\}_{t\geq 0} \subseteq \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ has a generator $Q \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ such that $N_t = e^{tQ}$. A classical result states that N_t is non-negative for all $t \geq 0$ if and only if the generator Q can be written in the form $Q = \Phi - K$, where Φ is non-negative and K is a diagonal matrix with respect to the canonical basis.[24] A second, crucial observation is that N_t is Heisenberg $B \not\to A$ semicausal for all $t \geq 0$

if and only if Q is Heisenberg $B \not\to A$ semicausal. To see this, let us first show that the reduced maps $\{N_t^A\}_{t\geq 0}$ also form a norm-continuous semigroup of non-negative maps. Since non-negativity is clear, we derive the semigroup properties ($N_0^A = \mathbb{1}_A$, $N_{t+s}^A = N_t^A N_s^A$, and continuity) from the corresponding ones of $\{N_t\}_{t\geq 0}$,

$$\begin{aligned}
N_0^A &= (\mathbb{1}_A \otimes \langle b_1|)(N_0^A \otimes |\mathbf{1}_B\rangle) = (\mathbb{1}_A \otimes \langle b_1|)N_0(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) = (\mathbb{1}_A \otimes \langle b_1|)(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) = \mathbb{1}_A,\\
N_{t+s}^A &= (\mathbb{1}_A \otimes \langle b_1|)(N_{t+s}^A \otimes |\mathbf{1}_B\rangle) = (\mathbb{1}_A \otimes \langle b_1|)N_{t+s}(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) = (\mathbb{1}_A \otimes \langle b_1|)N_tN_s(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)\\
&= (\mathbb{1}_A \otimes \langle b_1|)N_t(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)N_s^A = (\mathbb{1}_A \otimes \langle b_1|)(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)N_t^AN_s^A = N_t^AN_s^A,\\
\|N_t^A - N_s^A\| &= \sup_{\|x\|_\infty=1} \|(N_t^A - N_s^A)|x\rangle\|_\infty = \sup_{\|x\|_\infty=1} \|((N_t^A - N_s^A)|x\rangle) \otimes |\mathbf{1}_B\rangle\|_\infty = \sup_{\|x\|_\infty=1} \|(N_t - N_s)(|x\rangle \otimes |\mathbf{1}_B\rangle)\|_\infty\\
&\leq \sup_{\|y\|_\infty=1} \|(N_t - N_s)|y\rangle\| = \|N_t - N_s\|.
\end{aligned}$$

Thus, we conclude that $N_t^A = e^{tQ^A}$ for some generator $Q^A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$. We further have

$$\begin{aligned}
Q(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) &= \frac{d}{dt}\Big|_{t=0} N_t(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)\\
&= \frac{d}{dt}\Big|_{t=0} (\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)N_t^A\\
&= (\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)Q^A.
\end{aligned}$$

Thus, Q is semicausal if N_t is semicausal for all $t \geq 0$. Conversely, if Q is semicausal, then N_t is semicausal, since

$$\begin{aligned}
N_t(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) &= e^{tQ}(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)\\
&= \sum_{k=0}^{\infty} \frac{t^k}{k!} Q^k(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)\\
&= \sum_{k=0}^{\infty} \frac{t^k}{k!} (\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)\left(Q^A\right)^k\\
&= (\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)e^{tQ^A}.
\end{aligned}$$

Therefore, our task reduces to characterizing semicausal maps of the form $Q = \Phi - K$. Let us first remark that it is straight-forward to check (numerically) whether a given map satisfies these two conditions: We just need to check for non-negativity of the off-diagonal elements and whether $(\mathbb{1}_A \otimes \langle b|)Q|a_i \otimes \mathbf{1}_B\rangle = 0$ for all $a_i \in \mathbb{A}$ and all $b \in \{|\mathbf{1}_B\rangle\}^\perp$. That is, semicausality can be checked in terms of $|\mathbb{A}|(|\mathbb{B}| - 1)$ linear equations and $|\mathbb{A}||\mathbb{B}|(|\mathbb{A}||\mathbb{B}| - 1)$ linear inequalities. Thus, a desirable result would be a normal form for all Heisenberg $B \not\to A$ semicausal generators Q, which allows for generating such maps rather than checking whether a given maps is of the desired form. The main result of this section is exactly such a normal form.

To understand our normal form below, note that there are two natural ways of constructing a generator (remember that the matrix elements are interpreted as transition rates) that does not transmit information from system B to system A. First, we can leave system A unchanged and have transitions only on system B. The most basic form of such a map is $|a_i\rangle\langle a_i| \otimes B^{(i)}$ for some $1 \leq i \leq |\mathbb{A}|$ and for some $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ that is itself a valid generator of a semigroup of row-stochastic maps. That means that $B^{(i)} = \Phi^{(i)} - \mathrm{diag}(\Phi^{(i)}|\mathbf{1}_B\rangle)$ for some non-negative matrix $\Phi^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$. Second, if we want to act non-trivially on system A, we can make both the two parts of a generator $Q = \Phi - K$, the non-negative part $\Phi \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ and the diagonal part $K \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$, semicausal separately. Such a map has the form $\Phi_{sc} - K_A \otimes \mathbb{1}_B$, where Φ_{sc} is semicausal non-negative and $K_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$ is diagonal. The fact that (convex) combinations of these basic building blocks already give rise to the most general form of semicausal generators for semigroups of non-negative bounded linear maps is the content of our next theorem, which establishes the desired normal form.

Theorem IV.7 (generators of classical semigroups of semicausal non-negative maps). *A map $Q \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is the generator of a (norm-continuous) semigroup of Heisenberg $B \not\to A$ semicausal non-negative linear maps if and only if there exist a non-negative Heisenberg $B \not\to A$ semicausal map $\Phi_{sc} \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$, a diagonal map $K_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$, and linear maps $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ that generate (norm-continuous) semigroups of row-stochastic maps, for $1 \leq i \leq |\mathbb{A}|$, such that*

$$Q = \Phi_{sc} - K_A \otimes \mathbb{1}_B + \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes B^{(i)}.$$

In that case, Φ_{sc} can be chosen "block-off-diagonal," i.e., $\Phi_{sc} = \sum_{i\neq j} |a_i\rangle\langle a_j| \otimes \Phi_{sc}^{(ij)}$ for some collection of (non-negative) maps $\Phi_{sc}^{(ij)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$.

Proof. It is straight-forward to check that a generator Q of the given form has non-negative off-diagonal entries with respect to the standard basis and is Heisenberg $B \not\to A$ semicausal. By the above discussion, this means that such a generator indeed gives rise to a semigroup of semicausal non-negative maps.

We prove the converse. Suppose Q is the generator of a semigroup of non-negative linear maps. Then, we can expand it as $Q = \sum_{i,\, j=1}^{|\mathbb{A}|} |a_i\rangle\langle a_j| \otimes Q^{(ij)}$, where the operators $Q^{(ij)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ are non-negative for $i \neq j$ and of the form of a generator of a non-negative semigroup (i.e., non-negative minus diagonal) for $i = j$. This decomposition, together with semicausality, implies that for all $1 \leq i,j \leq |\mathbb{A}|$,

$$Q^{(ij)}|\mathbf{1}_B\rangle = (\langle a_i| \otimes \mathbf{1}_B)Q(|a_j\rangle \otimes |\mathbf{1}_B\rangle) = \langle a_i|Q^A|a_j\rangle \cdot |\mathbf{1}_B\rangle.$$

In other words, $|\mathbf{1}_B\rangle$ is an eigenvector of every $Q^{(ij)}$, with the corresponding eigenvalue $\lambda^{(ij)} = \langle a_i|Q^A|a_j\rangle$. Hence, if we define $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ as $B^{(i)} := Q^{(ii)} - \lambda^{(ii)}\mathbb{1}_B$, then B^i generates a semigroup of non-negative maps (since $Q^{(ij)}$ does and $\lambda^{(ii)}\mathbb{1}_B$ is diagonal) and satisfies (by construction) $B^{(i)}|\mathbf{1}_B\rangle = 0$. Hence, $B^{(i)}$ generates a semigroup of row-stochastic maps.

With this notation, we can rewrite Q as

$$Q = \underbrace{\sum_{i\neq j} |a_i\rangle\langle a_j| \otimes Q^{(ij)}}_{=:\Phi_{sc}} \underbrace{- \sum_{i=1}^{|\mathbb{A}|} -\lambda^{(ii)}|a_i\rangle\langle a_i|}_{=:K_A} \otimes \mathbb{1}_B + \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes B^{(i)}.$$

Note that Φ_{sc} is semicausal, since it can be written as the linear combination of the three semicausal maps Q, $K_A \otimes \mathbb{1}_B$, and $\sum_i |a_i\rangle\langle a_i| \otimes B^{(i)}$. Thus, we have reached the claimed form. □

By applying Theorem IV.4, we can further expand the Φ part.

Corollary IV.8. A map $Q \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is the generator of a (norm-continuous) semigroup of Heisenberg $B \not\to A$ semicausal non-negative linear maps if and only if there exist a (finite) alphabet $\mathbb{E}$, a (non-negative) row-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{E}}; \mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}})$, a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}}; \mathbb{R}^{\mathbb{A}})$, a diagonal matrix $K_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$, and maps $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ that generate (norm-continuous) semigroups of (row-)stochastic maps, for $1 \leq i \leq |\mathbb{A}|$, such that

$$Q = (A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes U) - K_A \otimes \mathbb{1}_B + \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes B^{(i)}.$$

In that case, we can choose $|\mathbb{E}| = |\mathbb{A}|^2$.

One should also note that with the notation of Corollary IV.8, the reduced map is given by $Q^A = (A(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle)) - K_A$. Hence, the reduced dynamics only depends on the operators A and K_A. Further note that if we require the semigroup to consist of non-negative semicausal maps that are also row-stochastic, then we obtain the additional requirement that $K_A|\mathbf{1}_A\rangle = A|\mathbf{1}_{AE}\rangle$, which completely determines K_A. For completeness and later use, we write down the form of the generators non-negative semigroups that are Schrödinger $B \not\to A$ semicausal.

Corollary IV.9. A map $Q \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$ is the generator of a (norm-continuous) semigroup of Schrödinger $B \not\to A$ semicausal non-negative linear maps if and only if there exist a (finite) alphabet $\mathbb{E}$, a (non-negative) column-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{B}})$, a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}})$, a diagonal matrix $K_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{B}})$, and maps $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ that generate (norm-continuous) semigroups of column-stochastic maps, for $1 \leq i \leq |\mathbb{A}|$, such that

$$Q = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B) - K_A \otimes \mathbb{1}_B + \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes B^{(i)}.$$

In that case, we can choose $|\mathbb{E}| = |\mathbb{A}|^2$.

Similar to the row-stochastic case, $B^{(i)}$ generates a semigroup of column-stochastic maps if and only if $B^{(i)} = \Phi^{(i)} - \mathrm{diag}(\langle \mathbf{1}_B|\Phi^{(i)})$ for some non-negative matrix $\Phi^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$.

D. Generators of semigroups of classical superchannels

We finally turn to semigroups of classical superchannels, that is, a collection of classical superchannels $\{\hat{S}_t\}_{t\geq 0}$, such that $\hat{S}_0 = \mathrm{id}$, $\hat{S}_{t+s} = \hat{S}_t\hat{S}_s$, and the map $t \mapsto \hat{S}_t$ is continuous (with respect to any and, thus, all of the equivalent norms in finite dimensions). To formulate a technically slightly stronger result, we call a linear map $\hat{S}$ a preselecting supermap if $\mathfrak{C}^C_{A;B} \circ \hat{S} \circ (\mathfrak{C}^C_{A;B})^{-1}$ is a non-negative Schrödinger

$B \not\to A$ semicausal map. Theorem IV.3 then tells us that a superchannel is a special preselecting supermap. The result of this section is the following theorem:

Theorem IV.10. *A linear map $\hat{Q} : \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}}) \to \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}})$ generates a semigroup of classical preselecting supermaps if and only if there exist a (finite) alphabet $\mathbb{E}$, a column-stochastic matrix $U \in \mathcal{B}(\mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{B}})$, a non-negative matrix $A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}})$, a diagonal matrix $K_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$, and a collection of generators of semigroups of column-stochastic matrices $B^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ such that*

$$\hat{Q}(M) = U(M \otimes \mathbb{1}_E)A - MK_A + \sum_{i=1}^{|\mathbb{A}|} B^{(i)} M |a_i\rangle\langle a_i|. \tag{5}$$

Furthermore, $\hat{Q}$ generates a semigroup of classical superchannels if and only if $\hat{Q}$ generates a semigroup of preselecting supermaps and $\langle a_i|K_A a_i\rangle = \langle \mathbf{1}_{AE}|Aa_i\rangle$ for all $1 \le i \le |\mathbb{A}|$. In this case, $\hat{Q}$ is given by

$$\hat{Q}(M) = U(M \otimes \mathbb{1}_E)A - \sum_{i=1}^{|\mathbb{A}|} \langle \mathbf{1}_{AE}|Aa_i\rangle M|a_i\rangle\langle a_i| + \sum_{i=1}^{|\mathbb{A}|} B^{(i)} M|a_i\rangle\langle a_i|. \tag{6}$$

Proof. The main idea is to relate the generators of superchannels to those of semicausal maps. This relation is given by definition for preselecting supermaps and by Theorem IV.3 for superchannels. For a generator $\hat{Q}$ of a semigroup of preselecting supermaps $\{\hat{S}_t\}_{t\ge 0}$, we have

$$\hat{Q} = \frac{d}{dt}\Big|_{t=0} \hat{S}_t = (\mathfrak{C}^C_{A;B})^{-1} \frac{d}{dt}\Big|_{t=0} \left[\mathfrak{C}^C_{A;B} \circ \hat{S}_t \circ (\mathfrak{C}^C_{A;B})^{-1}\right] \mathfrak{C}^C_{A;B}.$$

Thus, $\hat{Q}$ generates a semigroup of preselecting supermaps if and only if $\hat{Q}$ can be written as $\hat{Q} = (\mathfrak{C}^C_{A;B})^{-1} \circ Q \circ \mathfrak{C}^C_{A;B}$ for some generator Q of a semigroup of non-negative Schrödinger $B \not\to A$ semicausal maps. Thus, to prove the first part of our theorem, we simply take the normal form in Corollary IV.9 and compute the similarity transformation above.

For $|\Omega\rangle = \sum_i |a_i\rangle \otimes |a_i\rangle \in \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{A}}$ and an operator $X_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$, the well-known identity $(X_A \otimes \mathbb{1}_A)|\Omega\rangle = (\mathbb{1}_A \otimes X_A^T)|\Omega\rangle$ can be proven by a direct calculation. Similarly, it is easy to show that for $X_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}})$, the slightly more general identity $(X_A \otimes \mathbb{1}_A)|\Omega\rangle = (\mathbb{1}_A \otimes \mathbb{F}_{A;E} X_A^{T_A})|\Omega\rangle$ holds, where $\mathbb{F}_{A;E}$ is the flip operator that exchanges systems A and E. We use these two identities in the following calculations.

For $\tilde{A} \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{A}} \otimes \mathbb{R}^{\mathbb{E}})$ and $\tilde{U} \in \mathcal{B}(\mathbb{R}^{\mathbb{E}} \otimes \mathbb{R}^{\mathbb{B}}; \mathbb{R}^{\mathbb{B}})$, we have, for any $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}})$,

$$\begin{aligned}
(\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes \tilde{U})(\tilde{A} \otimes \mathbb{1}_B)\mathfrak{C}^C_{A;B}(M) &= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes \tilde{U})(\tilde{A} \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes M)|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes (\tilde{U}(\mathbb{1}_E \otimes M)))(\tilde{A} \otimes \mathbb{1}_A)|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes (\tilde{U}(\mathbb{1}_E \otimes M)))(\mathbb{1}_A \otimes \mathbb{F}_{A;E}\tilde{A}^{T_A})|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes (\tilde{U}(\mathbb{1}_E \otimes M)\mathbb{F}_{A;E}\tilde{A}^{T_A}))|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}\mathfrak{C}^C_{A;B}(\tilde{U}(\mathbb{1}_E \otimes M)\mathbb{F}_{A;E}\tilde{A}^{T_A}) \\
&= \tilde{U}(\mathbb{1}_E \otimes M)\mathbb{F}_{A;E}\tilde{A}^{T_A} \\
&= (\tilde{U}\mathbb{F}_{B;E})(M \otimes \mathbb{1}_E)\tilde{A}^{T_A}.
\end{aligned}$$

For $\tilde{K}_A \in \mathcal{B}(\mathbb{R}^{\mathbb{A}})$, we get, for any $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}})$,

$$\begin{aligned}
(\mathfrak{C}^C_{A;B})^{-1}(K_A \otimes \mathbb{1}_B)\mathfrak{C}^C_{A;B}(M) &= (\mathfrak{C}^C_{A;B})^{-1}(\tilde{K}_A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes M)|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes M)(\tilde{K}_A \otimes \mathbb{1}_A)|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes M)(\mathbb{1}_A \otimes \tilde{K}_A^T)|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A \otimes M\tilde{K}_A^T)|\Omega\rangle \\
&= (\mathfrak{C}^C_{A;B})^{-1}\mathfrak{C}^C_{A;B}(M\tilde{K}_A^T) \\
&= M\tilde{K}_A^T.
\end{aligned}$$

Finally, for an operator $\tilde{B}^{(i)} \in \mathcal{B}(\mathbb{R}^{\mathbb{B}})$ and for any $1 \le i \le |\mathbb{A}|$, we have, for any $M \in \mathcal{B}(\mathbb{R}^{\mathbb{A}}; \mathbb{R}^{\mathbb{B}})$,

$$
\begin{aligned}
(\mathfrak{C}^C_{A;B})^{-1}(|a_i\rangle\langle a_i|\otimes B^{(i)})\mathfrak{C}^C_{A;B}(M) &= (\mathfrak{C}^C_{A;B})^{-1}(|a_i\rangle\langle a_i|\otimes B^{(i)})(\mathbb{1}_A\otimes M)|\Omega\rangle\\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A\otimes B^{(i)}M)(|a_i\rangle\langle a_i|\otimes\mathbb{1}_A)|\Omega\rangle\\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A\otimes B^{(i)}M)(\mathbb{1}_A\otimes|a_i\rangle\langle a_i|)|\Omega\rangle\\
&= (\mathfrak{C}^C_{A;B})^{-1}(\mathbb{1}_A\otimes B^{(i)}M|a_i\rangle\langle a_i|)|\Omega\rangle\\
&= (\mathfrak{C}^C_{A;B})^{-1}\mathfrak{C}^C_{A;B}(B^{(i)}M|a_i\rangle\langle a_i|)\\
&= B^{(i)}M|a_i\rangle\langle a_i|.
\end{aligned}
$$

Applying the results of these calculations term by term to the normal form in Corollary IV.9 yields the first claim, where we defined $A = \tilde{A}^{T_A}$, $U = \tilde{U}\mathbb{F}_{B;E}$, $K_A = \tilde{K}_A^T$, and $B^{(i)} = \tilde{B}^{(i)}$.

If the semigroup $\{\hat{S}_t\}_{t\geq 0}$ consists of superchannels, that is, preselecting maps such that (by Theorem IV.3) the reduced maps S_t^A of the semigroup of semicausal maps $S_t := \mathfrak{C}^C_{A;B}\circ\hat{S}_t\circ(\mathfrak{C}^C_{A;B})^{-1}$ (which are defined by the requirement that $(\mathbb{1}_A\otimes|\mathbf{1}_B\rangle)S_t = S_t^A(\mathbb{1}_A\otimes|\mathbf{1}_B\rangle)$) satisfy $S_t^A|\mathbf{1}_A\rangle = |\mathbf{1}_A\rangle$, then differentiating this relation yields

$$
Q^A|\mathbf{1}_A\rangle = \left.\frac{d}{dt}\right|_{t=0} S_t^A|\mathbf{1}_A\rangle = \left.\frac{d}{dt}\right|_{t=0}|\mathbf{1}_A\rangle = 0.
$$

We conclude that $\hat{Q}$ generates a semigroup of superchannels if and only if Q generates a semigroup of semicausal maps and $Q^A|\mathbf{1}_A\rangle = 0$. We obtain directly from Corollary IV.9 that $Q^A = (\mathbb{1}_A\otimes|\mathbf{1}_E\rangle)\tilde{A} - \tilde{K}_A$. It follows that

$$
\langle a_i\mathbf{1}_E|\tilde{A}\mathbf{1}_A\rangle = \langle a_i\mathbf{1}_E|A^{T_A}\mathbf{1}_A\rangle = \langle\mathbf{1}_{AE}|Aa_i\rangle = \langle a_i|\tilde{K}_A\mathbf{1}_A\rangle = \langle a_i|K_Aa_i\rangle, \tag{7}
$$

where we used that $\tilde{K}_A = K_A$ is diagonal in the last step. This is the condition claimed in the theorem. Finally, (6) is obtained by combining this condition with (5). □

V. THE QUANTUM CASE

We now turn to the quantum case. As introduced and described in more detail in Ref. 1, a quantum superchannel is a map that maps quantum channels to quantum channels while preserving the probabilistic structure of the theory. To achieve the latter, it is usually required that a quantum superchannel is a linear map and that probabilistic transformations, i.e., trace non-increasing CP-maps, should be mapped to probabilistic transformations even if we add an innocent bystander. When dealing with superchannels, we will restrict ourselves to the finite-dimensional case and leave the infinite-dimensional case[25] for future work. We follow[1] and define superchannels as follows:

Definition V.1 (superchannels). A linear map $\hat{S}:\mathcal{B}(\mathcal{S}_1(\mathcal{H}_A);\mathcal{S}_1(\mathcal{H}_B))\to\mathcal{B}(\mathcal{S}_1(\mathcal{H}_A);\mathcal{S}_1(\mathcal{H}_B))$ is called a superchannel if for all $n\in\mathbb{N}$ the map $\hat{S}_n = \mathrm{id}_{\mathcal{B}(\mathcal{S}_1(\mathbb{C}^n))}\otimes\hat{S}$ satisfies that $\hat{S}_n(T)$ is a probabilistic transformation whenever $T\in\mathcal{B}(\mathcal{S}_1(\mathbb{C}^n\otimes\mathcal{H}_A);\mathcal{S}_1(\mathbb{C}^n\otimes\mathcal{H}_B))$ is a probabilistic transformation and that $\hat{S}_n(T)$ is a quantum channel whenever $T\in\mathcal{B}(\mathcal{S}_1(\mathbb{C}^n\otimes\mathcal{H}_A);\mathcal{S}_1(\mathbb{C}^n\otimes\mathcal{H}_B))$ is a quantum channel.

A related concept is that of a semicausal quantum channel, which is a quantum channel on a bipartite space $\mathcal{H}_A\otimes\mathcal{H}_B$ such that no communication from B to A is allowed. Following Refs 2 and 10, we formalize this as follows:

Definition V.2 (semicausality). A bounded linear map $L_*:\mathcal{S}_1(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{S}_1(\mathcal{H}_A\otimes\mathcal{H}_B)$ is called Schrödinger $B\not\to A$ semicausal if there exists $L_*^A:\mathcal{S}_1(\mathcal{H}_A)\to\mathcal{S}_1(\mathcal{H}_A)$ such that $\mathrm{tr}_B[L_*(\rho)] = L_*^A(\mathrm{tr}_B[\rho])$, for all $\rho\in\mathcal{S}_1(\mathcal{H}_A\otimes\mathcal{H}_B)$. Similarly, $L:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$ is called Heisenberg $B\not\to A$ semicausal if there exists $L^A:\mathcal{B}(\mathcal{H}_A)\to\mathcal{B}(\mathcal{H}_A)$ such that $L(X_A\otimes\mathbb{1}_B) = L^A(X_A)\otimes\mathbb{1}_B$ for all $X_A\in\mathcal{B}(\mathcal{H}_A)$.

The map L_* is Schrödinger $B\not\to A$ semicausal if and only if the dual map $L := L_*^*$ is normal and Heisenberg $B\not\to A$ semicausal. We will often omit the Schrödinger or Heisenberg attribute if it is clear from the context. This section is structured analogously to the section about the classical case. In particular, we will start by reminding the reader of the connection between semicausal maps and superchannels as well as the characterization of semicausal CP-maps in terms of semilocalizable maps, as schematically shown in Fig. 4. We then turn to the study of the generators of semigroups of semicausal CP-maps and finally use the correspondence between superchannels and semicausal channels to obtain the corresponding results of the generators of semigroups of superchannels.

A. Superchannels, semicausal channels, and semilocalizable channels

We first state the characterization of superchannels in terms of semicausal maps, obtained in Ref. 1.

Theorem V.3. *For finite-dimensional spaces $\mathcal{H}_A$ and $\mathcal{H}_B$, let $\hat{S}:\mathcal{B}(\mathcal{S}_1(\mathcal{H}_A);\mathcal{S}_1(\mathcal{H}_B))\to\mathcal{B}(\mathcal{S}_1(\mathcal{H}_A);\mathcal{S}_1(\mathcal{H}_B))$ be a linear map and define $S = \mathfrak{C}_{A;B}\circ\hat{S}\circ\mathfrak{C}^{-1}_{A;B}$. Then, $\hat{S}$ is a superchannel if and only if S is CP and Schrödinger $B\not\to A$ semicausal such that the reduced map S^A satisfies $S^A(\mathbb{1}_A) = \mathbb{1}_A$.*

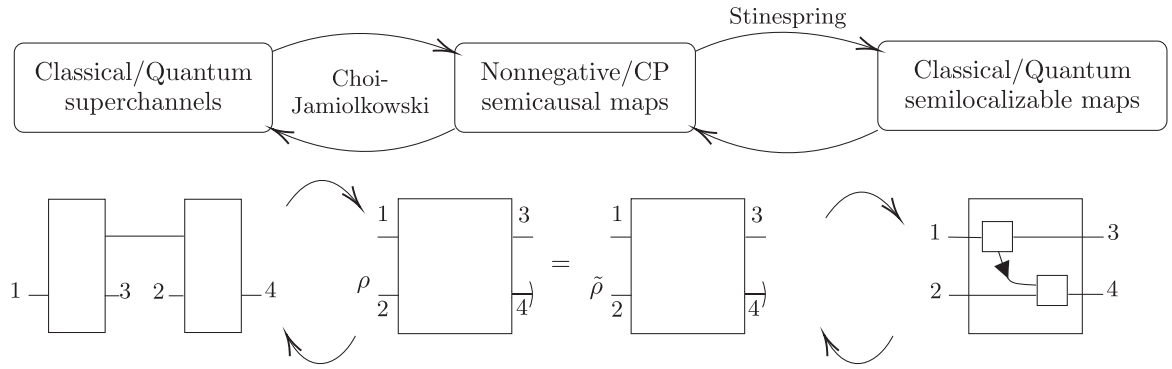


FIG. 4. Visualization of the relation between the notions of superchannels, semicausal maps, and semilocalizable maps. Superchannels and semicausal maps are related via a similarity transform with the Choi–Jamiołkowski isomorphism. Schrödinger $B \not\to A$ semicausal maps are those maps whose output, after tracing out system 4, does not depend on input 2 (ρ or $\tilde{\rho}$). Semicausal maps are precisely those maps that allow for one-way communication only. This is called semilocalizability.

The next result is due to Eggeling, Schlingemann, and Werner,[10] who proved it in the finite-dimensional setting. The following form, which is a generalization of Ref. 10 to the infinite-dimensional case and which has previously been shown in (Ref. 26, Theorem 4), can be obtained from our main result (Theorem V.6) by setting $K = 0$.

Theorem V.4. *A map $\Phi \in CP_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$ is Heisenberg $B \not\to A$ semicausal if and only if there exist a (separable) Hilbert space $\mathcal{H}_E$, a unitary operator $U \in \mathcal{U}(\mathcal{H}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$, and an arbitrary operator $A \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_E)$ such that*

$$\Phi(X) = V^\dagger (X \otimes \mathbb{1}_E) V, \text{ with } V = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B). \tag{8}$$

If $\mathcal{H}_A$ and $\mathcal{H}_B$ are finite-dimensional, with dimensions d_A and d_B, then $\mathcal{H}_E$ can be chosen such that $\dim(\mathcal{H}_E) \leq (d_A d_B)^2$.

We call a normal CP-map $\Phi \in \mathrm{CP}_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$ *semilocalizable* if its Stinespring dilation can be written in the form of Eq. (8). With that nomenclature, the above theorem is exactly the quantum analog of Theorem IV.4.

B. Generators of semigroups of semicausal CP maps

The main goal of this section is to establish a structure theorem for the generators of semigroups of semicausal CP-maps, the proof-structure of which is highlighted in Fig. 5. This is our main technical contribution. To get started, recall that a generator $L \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B))$ generates a norm-continuous semigroup $\{T_t\}_{t\geq 0} \subseteq \mathrm{CP}_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$ of CP-maps (i.e., $T_t = e^{tL}$) if and only if L can be written in GKLS-form, i.e., if and only if there exist $\Phi \in \mathrm{CP}_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$ and $K \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ such that

$$L(X) = \Phi(X) - K^\dagger X - XK, \quad X \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B). \tag{9}$$

As in the classical case, we continue by showing that T_t is Heisenberg $B \not\to A$ semicausal for all $t \geq 0$ if and only if L is Heisenberg $B \not\to A$ semicausal. We start by showing that the family of reduced maps $\{T_t^A\}_{t\geq 0}$ also forms a norm-continuous semigroup of normal CP-maps. That T_t^A is normal and CP follows, since for any density operator $\rho_B \in \mathcal{S}_1(\mathcal{H}_B)$, we have

$$T_t^A = \mathrm{tr}_{\rho_B} \circ T_t \circ D,$$

where $D \in \mathrm{CP}_\sigma(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_B)$ is defined by $D(X_A) = X_A \otimes \mathbb{1}_B$. Hence, T_t^A is a normal CP-map as composition of normal CP-maps. It remains to check the semigroup properties ($T_0^A = \mathrm{id}_A$, $T_{t+s}^A = T_t^A T_s^A$, and norm-continuity). We have

$$\begin{aligned}
T_0^A(X_A) &= \mathrm{tr}_{\rho_B}[T_0(X_A \otimes \mathbb{1}_B)] = \mathrm{tr}_{\rho_B}[X_A \otimes \mathbb{1}_B] = X_A,\\
T_{t+s}^A(X_A) &= \mathrm{tr}_{\rho_B}[T_{t+s}(X_A \otimes \mathbb{1}_B)] = \mathrm{tr}_{\rho_B}[T_t(T_s(X_A \otimes \mathbb{1}_B))] = \mathrm{tr}_{\rho_B}\big[T_t(T_s^A(X_A) \otimes \mathbb{1}_B)\big] = \mathrm{tr}_{\rho_B}\big[(T_t^A T_s^A(X_A)) \otimes \mathbb{1}_B\big] = T_t^A T_s^A(X_A),\\
\big\|T_t^A - T_s^A\big\| &= \sup_{\|X_A\|_{\mathcal{B}(\mathcal{H}_A)}=1} \big\|T_t^A(X_A) - T_s^A(X_A)\big\|_{\mathcal{B}(\mathcal{H}_A)} = \sup_{\|X_A\|_{\mathcal{B}(\mathcal{H}_A)}=1} \big\|(T_t^A(X_A) - T_s^A(X_A)) \otimes \mathbb{1}_B\big\|_{\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)}\\
&= \sup_{\|X_A\|_{\mathcal{B}(\mathcal{H}_A)}=1} \big\|T_t(X_A \otimes \mathbb{1}_B) - T_s(X_A \otimes \mathbb{1}_B)\big\|_{\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)} \leq \sup_{\|X\|_{\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)}=1} \big\|T_t(X) - T_s(X)\big\|_{\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)} = \|T_t - T_s\|.
\end{aligned}$$

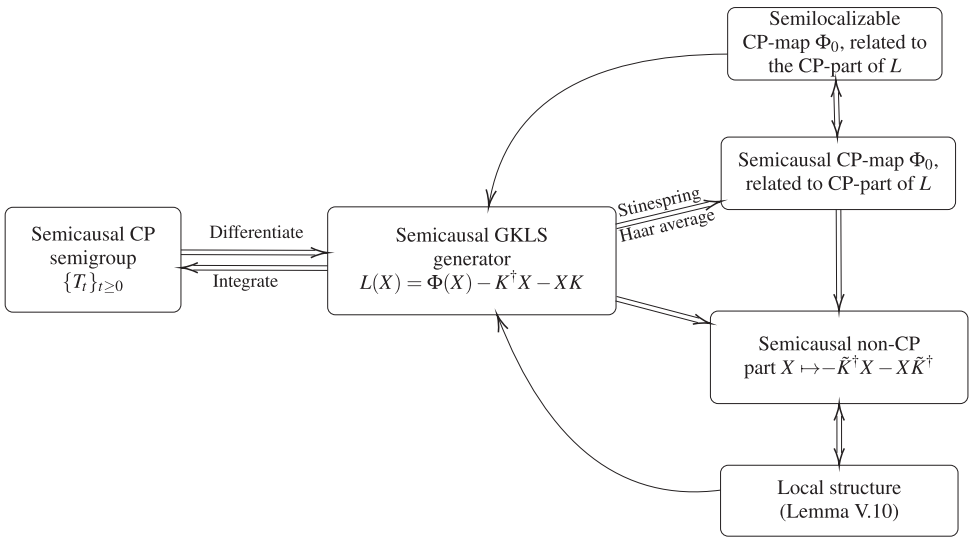


FIG. 5. Overview of the proof structure leading to the normal form for semicausal Lindblad generators (Theorem V.6). We first observe that semicausality of the CP semigroup is equivalent to semicausality of the corresponding GKLS generator L. The insight is then that we can construct a CP-map Φ_0 that is closely related to the CP-part of L and that is semicausal (Lemma V.13). From the semilocalizable form of Φ_0, we then obtain an explicit form for the CP-part of L. This, together with the observation that a semicausal non-CP part has to have a local form, yields the desired normal form.

Thus, we conclude that $T_t^A = e^{tL^A}$ for some generator $L^A \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A))$ of normal CP-maps. We further have

$$L(X_A \otimes \mathbb{1}_B) = \frac{d}{dt}\Big|_{t=0} T_t(X_A \otimes \mathbb{1}_B) = \frac{d}{dt}\Big|_{t=0} T_t^A(X_A) \otimes \mathbb{1}_B = L^A(X_A) \otimes \mathbb{1}_B.$$

Thus, L is semicausal if T_t is semicausal for all $t \geq 0$. Conversely, if L is semicausal, then T_t is semicausal for all $t \geq 0$, since

$$\begin{aligned} T_t(X_A \otimes \mathbb{1}_B) &= e^{tL}(X_A \otimes \mathbb{1}_B) \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!} L^k(X_A \otimes \mathbb{1}_B) \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \left(L^A\right)^k (X_A) \otimes \mathbb{1}_B \\ &= e^{tL^A}(X_A) \otimes \mathbb{1}_B. \end{aligned}$$

Therefore, our task reduces to characterizing semicausal maps in the GKLS-form, i.e., we want to determine the corresponding Φ and K. Our main result (Theorem V.6) is a normal form, which allows us to list all semicausal generators L.

Before we delve into this, we treat the inverse question: Given some $L \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B))$, is it a semicausal generator? A computationally efficiently chackable criterion can be constructed via the Choi–Jamiołkowski isomorphism. If $\mathcal{H}_A$ and $\mathcal{H}_B$ are finite-dimensional and $L \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B))$ is given, then we define $\mathfrak{L} = \mathfrak{C}_{AB;AB}(L) \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$, where the Choi–Jamiołkowski isomorphism is defined with respect to the orthogonal bases $\{|a_i\rangle\}_{i=1}^{\dim(\mathcal{H}_A)}$ and $\{|b_j\rangle\}_{j=1}^{\dim(\mathcal{H}_B)}$ of $\mathcal{H}_A$ and $\mathcal{H}_B$, respectively, and where the spaces $\mathcal{H}_{A_1} = \mathcal{H}_{A_2} = \mathcal{H}_A$ and $\mathcal{H}_{B_1} = \mathcal{H}_{B_2} = \mathcal{H}_B$ are introduced for notational convenience. Furthermore, define $P^\perp \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$ to be the orthogonal projection onto the orthogonal complement of $\{|\Omega\rangle\}$, where $|\Omega\rangle = \sum_{i,j} |a_i\rangle \otimes |b_j\rangle \otimes |a_i\rangle \otimes |b_j\rangle$.

Lemma V.5. A linear map $L : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ is the generator of a semigroup of Heisenberg $B \nrightarrow A$ semicausal CP-maps if and only if

- *$\mathfrak{L}$ is self-adjoint and $P^\perp \mathfrak{L} P^\perp \geq 0$, and*
- *$\mathrm{tr}_{B_1}[\mathfrak{L}] = \mathfrak{L}^A \otimes \mathbb{1}_{B_2}$ for some (then necessarily self-adjoint) $\mathfrak{L}^A \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{A_2})$.*

The generated semigroup is unital (i.e., $T_t(\mathbb{1}_{AB}) = \mathbb{1}_{AB}$ for $t \geq 0$) if and only if $\mathrm{tr}_{A_1}\left[\mathfrak{L}^A\right] = 0$.

Furthermore, a linear map $L : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ is the generator of a semigroup of Schrödinger $B \nrightarrow A$ semicausal CP-maps if and only if

- $\mathfrak{L}$ *is self-adjoint and* $P^{\perp}\mathfrak{L}P^{\perp} \geq 0$ *and*
- $(\mathbb{F}_{A_1;B_1} \otimes \mathbb{1}_{A_2})\mathrm{tr}_{B_2}[\mathfrak{L}](\mathbb{F}_{A_1;B_1} \otimes \mathbb{1}_{A_2}) = \mathbb{1}_{B_1} \otimes \mathfrak{L}^A$ *for some (then necessarily self-adjoint)* $\mathfrak{L}^A \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{A_2})$.

The generated semigroup is trace-preserving (i.e., $\mathrm{tr}[T_t(\rho)] = \mathrm{tr}[\rho]$ *for* $\rho \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ *and* $t \geq 0$*) if and only if* $\mathrm{tr}_{A_2}[\mathfrak{L}^A] = 0$.

Thus, checking whether a map L is the generator of a semigroup of semicausal CP-maps reduces to checking several semidefinite constraints. In particular, the problem to optimize over all semicausal generators is a semidefinite program.

Proof. It is known (see, e.g., the appendix in Ref. 21) that L generates a semigroup of CP-maps if and only if $\mathfrak{L}$ is self-adjoint and $P^{\perp}\mathfrak{L}P^{\perp} \geq 0$. This criterion goes by the name of conditional complete positivity.[22] Thus, it remains to translate the other criteria to the level of Choi–Jamiołkowski operators. If L is Heisenberg $B \not\to A$ semicausal, then

$$\begin{aligned}\mathrm{tr}_{B_1}[\mathfrak{L}] &= \mathrm{tr}_{B_1}[(\mathrm{id}_{A_1B_1} \otimes L)(|\Omega\rangle\langle\Omega|)]\\ &= (\mathrm{id}_{A_1} \otimes L)(|\Omega_A\rangle\langle\Omega_A| \otimes \mathbb{1}_{B_2})\\ &= (\mathrm{id}_{A_1} \otimes L^A)(|\Omega_A\rangle\langle\Omega_A|) \otimes \mathbb{1}_{B_2}\\ &= \mathfrak{L}^A \otimes \mathbb{1}_{B_2},\end{aligned}$$

where we defined $|\Omega_A\rangle = \sum_i |a_i\rangle \otimes |a_i\rangle \in \mathcal{H}_{A_1} \otimes \mathcal{H}_{A_2}$ and $\mathfrak{L}^A = (\mathrm{id}_{A_1} \otimes L^A)(|\Omega_A\rangle\langle\Omega_A|)$. Conversely, if $\mathrm{tr}_{B_1}[\mathfrak{L}] = \mathfrak{L}^A \otimes \mathbb{1}_{B_2}$, define $L^A = \mathfrak{C}^{-1}_{A;A}(\mathfrak{L}^A)$. Then,

$$\begin{aligned}L(X_A \otimes \mathbb{1}_{B_1}) &= \mathrm{tr}_{A_1B_1}\left[\left((X_A^T \otimes \mathbb{1}_{B_1}) \otimes \mathbb{1}_{A_2B_2}\right)\mathfrak{L}\right]\\ &= \mathrm{tr}_{A_1}\left[\left(X_A^T \otimes \mathbb{1}_{A_2B_2}\right)\mathrm{tr}_{B_1}[\mathfrak{L}]\right]\\ &= \mathrm{tr}_{A_1}\left[(X_A^T \otimes \mathbb{1}_{A_2B_2})(\mathfrak{L}^A \otimes \mathbb{1}_{B_2})\right]\\ &= \mathrm{tr}_{A_1}\left[(X_A^T \otimes \mathbb{1}_{A_2})\mathfrak{L}^A\right] \otimes \mathbb{1}_{B_2}\\ &= \mathfrak{C}^{-1}_{A;A}(\mathfrak{L}^A)(X_A) \otimes \mathbb{1}_{B_2}\\ &= L^A(X_A) \otimes \mathbb{1}_B.\end{aligned}$$

Finally, it is known that a semigroup of CP-maps is unital if and only if $L(\mathbb{1}_{A_2B_2}) = 0$. However, this is equivalent to our criterion, since a simple calculation shows that

$$\mathrm{tr}_{A_1B_1}[\mathfrak{L}] = L(\mathbb{1}_{A_2B_2}).$$

This finishes the proof for the Heisenberg picture case. The Schrödinger case can be proven along similar lines or be obtained directly from the Heisenberg case via the identity $\mathfrak{C}_{AB;AB}(L^*) = \mathbb{F}_{A_1B_1;A_2B_2}[\mathfrak{C}_{AB;AB}(L)]^T\mathbb{F}_{A_1B_1;A_2B_2}$. □

Let us now return to the main goal of this section: finding a normal form for semicausal generators in GKLS-form. We motivate (and interpret) our normal form as the "quantization" of the normal form for generators of classical semicausal semigroups (Theorem IV.7). In the classical case, the normal form had two building blocks: an operator of the form $Q_1 = \Phi_{sc} - K_A \otimes \mathbb{1}_B$, where Φ_{sc} is non-negative and semicausal, and an operator of the form $Q_2 = \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes B^{(i)}$, where the $B^{(i)}$'s are generators of row-stochastic maps, (i.e., $B^{(i)}$ generates a non-negative semigroup and $B^{(i)}|\mathbf{1}_B\rangle = 0$). It is straightforward to guess a quantum analog for the first building block: a generator $L_1 \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B))$ defined by

$$L_1(X) = \Phi_{sc}(X) - (K_A \otimes \mathbb{1}_B)^{\dagger}X - X(K_A \otimes \mathbb{1}_B), \tag{10}$$

where $\Phi_{sc} \in \mathrm{CP}_{\sigma}(\mathcal{H}_A \otimes \mathcal{H}_B)$, given in the Stinespring form by $\Phi_{sc}(X) = V_{sc}^{\dagger}(X \otimes \mathbb{1}_E)V_{sc}$, is semicausal. One readily verifies that L_1 defines a semicausal generator. To "quantize" the second building block, note that Q_2 does not induce any change on system A. Indeed, since

$$e^{tQ_2}(\mathbb{1}_A \otimes |\mathbf{1}_B\rangle) = \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes (e^{tB^{(i)}}|\mathbf{1}_B\rangle) = \sum_{i=1}^{|\mathbb{A}|} |a_i\rangle\langle a_i| \otimes |\mathbf{1}_B\rangle = \mathbb{1}_A \otimes |\mathbf{1}_B\rangle, \tag{11}$$

the generated semigroup looks like the identity on system A. In the quantum case, semigroups that do not induce any change on system A are more restricted, since any information-gain about system A inevitably disturbs system A—so there can be no conditioning as in the classical case. Indeed, if one requires that $T_t \in \mathrm{CP}_{\sigma}(\mathcal{H}_A \otimes \mathcal{H}_B)$ satisfies the quantum analog of Eq. (11), namely,

$$T_t(X_A \otimes \mathbb{1}_B) = X_A \otimes \mathbb{1}_B \tag{12}$$

for all $X_A \in \mathcal{B}(\mathcal{H}_A)$, then $T_t = \mathrm{id}_A \otimes \Theta_t$ for some unital map $\Theta_t \in \mathrm{CP}_\sigma(\mathcal{H}_B)$ (see Appendix B for a proof). Differentiation of $T_t = \mathrm{id}_A \otimes \Theta_t$ at $t = 0$ now implies that the generator of a semigroup of CP-maps that satisfy (12) are of the form $\mathrm{id}_A \otimes \hat{B}$, where $\hat{B}$ generates a semigroup of unital CP-maps [i.e., $\hat{B}(\mathbb{1}_B) = 0$]. To conclude, the two building blocks are operators of the form of L_1 in Eq. (10) and maps L_2 of the form

$$L_2(X) = (\mathbb{1}_A \otimes B)^\dagger (X \otimes \mathbb{1}_E)(\mathbb{1}_A \otimes B) - \frac{1}{2}\left\{\mathbb{1}_A \otimes B^\dagger B, X\right\} + i[\mathbb{1}_A \otimes H_B, X],$$

with $B \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$ and a self-adjoint $H_B \in \mathcal{B}(\mathcal{H}_B)$.

In the classical case, we obtained the normal form (Theorem IV.7) by taking a convex combination of the basic building blocks. This corresponds to probabilistically choosing one or the other. In quantum theory, there is a more general concept: superposition. To account for this, we construct our normal form not as a convex combination of the maps L_1 and L_2 but by taking a linear combination (superposition) of the Stinespring operators V_{sc} and $\mathbb{1}_A \otimes B$ as the Stinespring operator of the CP-part of the GKLS-form (note here that the coefficients can be absorbed into V_{sc} and $\mathbb{1}_A \otimes B$, respectively). This means that if L is given by Eq. (9) with $\Phi(X) = V^\dagger(X \otimes \mathbb{1}_E)V$, then we take $V = V_{sc} + \mathbb{1}_A \otimes B$. It turns out that K can then be chosen such that L becomes semicausal. Also note that we can further decompose $V_{sc} = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$, as in Theorem V.4.

Our main technical result is that the heuristics employed in the "quantization" procedure above is sound, i.e., that the generators constructed in that way are the only semicausal generators in the GKLS-form.

Theorem V.6. *Let $L : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ be defined by $L(X) = \Phi(X) - K^\dagger X - XK$, with $\Phi \in \mathrm{CP}_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$ and $K \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$. Then, L is Heisenberg $B \not\to A$ semicausal if and only if there exist a (separable) Hilbert space $\mathcal{H}_E$, a unitary $U \in \mathcal{U}(\mathcal{H}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$, a self-adjoint operator $H_B \in \mathcal{B}(\mathcal{H}_B)$, and arbitrary operators $A \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_E)$, $B \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$, and $K_A \in \mathcal{B}(\mathcal{H}_A)$ such that*

$$\Phi(X) = V^\dagger(X \otimes \mathbb{1}_E)V, \text{ with } V = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B) + (\mathbb{1}_A \otimes B), \tag{13a}$$

$$K = (\mathbb{1}_A \otimes B^\dagger U)(A \otimes \mathbb{1}_B) + \frac{1}{2}\mathbb{1}_A \otimes B^\dagger B + K_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes iH_B. \tag{13b}$$

If $\mathcal{H}_A$ and $\mathcal{H}_B$ are finite-dimensional, with dimensions d_A and d_B, then $\mathcal{H}_E$ can be chosen such that $\dim(\mathcal{H}_E) \leq (d_A d_B)^2$.

Remark V.7. Note that the characterization in Theorem V.6 is for generators of Heisenberg $B \not\to A$ semicausal dynamical semigroups. There are two special cases of interest: First, if we want the dynamical semigroup to be unital, then we need to further impose $L(\mathbb{1}_A \otimes \mathbb{1}_B) = 0$ in the normal form above, which is equivalent to $A^\dagger A = K_A + K_A^\dagger$—a constraint that also appears in the usual Linblad form. Second, if the dynamical semigroup corresponds (in the sense of Theorem V.3) to a semigroup of superchannels, then we additionally require that the reduced generator satisfies $L_*^A(\mathbb{1}_A) = 0$. We will use this in the "translation step" in Theorem V.18.

Remark V.8. In the finite-dimensional case, the Proof of Theorem V.6 is constructive. In Appendix C, we discuss in detail how to obtain the operators A, U, K_A, B, and H_B starting from the conditions in Lemma V.5.

The remainder of this section is devoted to the Proof of Theorem V.6, whose structure is highlighted in Fig. 5.

We begin with a technical observation about certain Haar integrals.

Lemma V.9. Let $\mathcal{H}_n$ be an n-dimensional subspace of $\mathcal{H}_A$ with orthogonal projection $P_n \in \mathcal{B}(\mathcal{H}_A)$, and let $V \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B; \mathcal{H}_A \otimes \mathcal{H}_C)$. Then,

$$\int_{\mathcal{U}_P(\mathcal{H}_n)} (U \otimes \mathbb{1}_C)\, V\, (U^\dagger \otimes \mathbb{1}_B)\, dU = P_n \otimes \frac{1}{n}\mathrm{tr}_{P_n}[V], \tag{14}$$

where the integration is with respect to the Haar measure on $\mathcal{U}_P(\mathcal{H}_n)$. It follows that $\left\|P_n \otimes \frac{1}{n}\mathrm{tr}_{P_n}[V]\right\| \leq \|V\|$. Furthermore, if $\mathcal{H}$ is separable infinite-dimensional, with orthonormal basis $\{|e_i\rangle\}_{i\in\mathbb{N}}$ and $\mathcal{H}_n = \mathrm{span}\{|e_1\rangle, |e_2\rangle, \ldots, |e_n\rangle\}$, then there exist $B \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_C)$ and an ultraweakly convergent subsequence of $\left(P_n \otimes \frac{1}{n}\mathrm{tr}_{P_n}[V]\right)_{n\in\mathbb{N}}$ with the limit $\mathbb{1}_A \otimes B$.

Proof. To calculate the integral, we employ the Weingarten formula,[27–29] which for the relevant case reads

$$\int_{\mathcal{U}_P(\mathcal{H}_n)} U_{ij} U^\dagger_{j',i'}\, dU = \frac{1}{n}\delta_{ii'}\delta_{jj'},$$

where $U_{ij} = \langle f_i | U f_j \rangle$ and $U^\dagger_{j'\,i'} = \langle f_{j'} | U^\dagger f_{i'} \rangle$ for some orthonormal basis $\{|f_1\rangle, |f_2\rangle, \ldots, |f_n\rangle\}$ of $\mathcal{H}_n$. A basis expansion then yields

$$\int_{\mathcal{U}_P(\mathcal{H}_n)} (U \otimes \mathbb{1}_C)\, V\, (U^\dagger \otimes \mathbb{1}_B)\, dU = \sum_{i,j,i',j'=1}^{n} \left[|f_i\rangle\langle f_{i'}| \otimes \left((\langle f_j| \otimes \mathbb{1}_C)\, V \left(|f_{j'}\rangle \otimes \mathbb{1}_B \right) \right) \int_{\mathcal{U}_P(\mathcal{H}_n)} U_{ij} U^\dagger_{j'\,i'}\, dU \right] = P_n \otimes \frac{1}{n} \mathrm{tr}_{P_n}[V].$$

For the second claim, we note that a standard estimate of the integral yields $\left\| \frac{1}{n} \mathrm{tr}_{P_n}[V] \right\| = \left\| P_n \otimes \frac{1}{n} \mathrm{tr}_{P_n}[V] \right\| \leq \|V\|$. Thus, the sequence $\left(\frac{1}{n} \mathrm{tr}_{P_n}[V] \right)_{n \in \mathbb{N}}$ is bounded and hence, by Banach–Alaoglu, has an ultraweakly convergent subsequence, whose limit we call B. The claim then follows by observing that under the separability assumption, $(P_n)_{n \in \mathbb{N}}$ converges ultraweakly to $\mathbb{1}_A$ and that the tensor product of two ultraweakly convergent sequences converges ultraweakly. □

As a first step toward our main result, we provide a characterization of those semicausal Lindblad generators that can be written with the vanishing CP part.

Lemma V.10. Let $L : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$, $L(X) := -K^\dagger X - XK$, with $K \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$. Then, L is Heisenberg $B \not\to A$ semicausal if and only if there exist $K_A \in \mathcal{B}(\mathcal{H}_A)$ and a self-adjoint $H_B \in \mathcal{B}(\mathcal{H}_B)$, with $K = K_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes iH_B$.

Proof. If $K = K_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes iH_B$, then $L(X_A \otimes \mathbb{1}_B) = (-K_A^\dagger X_A - K_A X_A) \otimes \mathbb{1}_B + X_A \otimes (iH_B - iH_B) = (-K_A^\dagger X_A - X_A K_A) \otimes \mathbb{1}_B$. Hence, L is semicausal. Conversely, suppose L is semicausal with $L(X_A \otimes \mathbb{1}_B) = L^A(X_A) \otimes \mathbb{1}_B$. Let $\mathcal{H}_n$ be an n-dimensional subspace of $\mathcal{H}_A$ and $U \in \mathcal{U}_P(\mathcal{H}_n)$. Then,

$$(L(U \otimes \mathbb{1}_B))(U^\dagger \otimes \mathbb{1}_B) = -K^\dagger (P_n \otimes \mathbb{1}_B) - (U \otimes \mathbb{1}_B) K (U^\dagger \otimes \mathbb{1}_B) = (L^A(U)\, U^\dagger) \otimes \mathbb{1}_B,$$

where $P_n \in \mathcal{B}(\mathcal{H}_A)$ is the orthogonal projection onto $\mathcal{H}_n$. We integrate both sides with respect to the Haar measure on $\mathcal{U}_P(\mathcal{H}_n)$. Lemma V.9 and some rearrangement and taking the conjugate yields

$$(P_n \otimes \mathbb{1}_B) K = -P_n \otimes \frac{1}{n} \mathrm{tr}_{P_n}\left[K^\dagger \right] - L_n^A \otimes \mathbb{1}_B \tag{15}$$

for some operator $L_n^A \in \mathcal{B}(\mathcal{H}_A)$. If $\mathcal{H}_A$ is finite-dimensional, we can take $\mathcal{H}_n = \mathcal{H}_A$ so that $P_n = \mathbb{1}_A$. Hence, $K = -\tilde{K}_A \otimes \mathbb{1}_B - \mathbb{1}_A \otimes B$, with $B = \frac{1}{n} \mathrm{tr}_A\left[K^\dagger \right]$ and $\tilde{K}_A = L_n^A$. If $\mathcal{H}_A$ is separable infinite-dimensional, we obtain the same result via a limiting procedure $n \to \infty$ as follows: Let $\{|e_i\rangle\}_{i \in \mathbb{N}}$ be an orthonormal basis of $\mathcal{H}_A$ and set $\mathcal{H}_n = \mathrm{span}\{|e_1\rangle, |e_2\rangle, \ldots, |e_n\rangle\}$. Then, the second part of Lemma V.9 allows us to pass to a subsequence of $\left(P_n \otimes \frac{1}{n} \mathrm{tr}_{P_n}\left[K^\dagger \right] \right)_{n \in \mathbb{N}}$ that converges ultraweakly to a limit $\mathbb{1}_A \otimes B$. The corresponding subsequence of $((P_n \otimes \mathbb{1}_B) K)_{n \in N}$ converges ultraweakly to K, and hence, that subsequence of $\left(L_n^A \otimes \mathbb{1}_B \right)_{n \in \mathbb{N}}$ converges ultraweakly to a limit $\tilde{K}_A \otimes \mathbb{1}_B$. That is, we get $K = -\tilde{K}_A \otimes \mathbb{1}_B - \mathbb{1}_A \otimes B$. Therefore,

$$0 = L(X_A \otimes \mathbb{1}_B) - L(X_A \otimes \mathbb{1}_B) = (L^A(X_A) - \tilde{K}_A^\dagger X_A - X_A \tilde{K}_A) \otimes \mathbb{1}_B - X_A \otimes (B + B^\dagger),$$

which can only be true for all X_A if $B + B^\dagger$ is proportional to $\mathbb{1}_B$. Since $B + B^\dagger$ is self-adjoint, we have $B + B^\dagger = 2r\mathbb{1}_B$ for some $r \in \mathbb{R}$. We can then set $iH_B := r\mathbb{1}_B - B$ and $K_A := -\tilde{K}_A - r\mathbb{1}$ so that H_B is self-adjoint and $K = K_A \otimes \mathbb{1} + \mathbb{1} \otimes iH_B$. □

If we had restricted our attention to Hamiltonian generators and unitary groups in finite dimensions, an analog of this lemma would have already followed from the fact that semicausal unitaries are tensor products, which was proved in Ref. 2 (and reproved in Ref. 11).

As another technical ingredient, the following lemma establishes a closedness property of the set of semicausal maps:

Lemma V.11. Let $(V_m)_{m \in \mathbb{N}}$ and $(W_n)_{n \in \mathbb{N}}$ be ultraweakly convergent sequences in $\mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B; \mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_E)$, with limits V and W. Suppose that for all $m, n \in \mathbb{N}$, the map $\Phi_{m,n} : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$, defined by $\Phi_{m,n}(X) = V_m^\dagger (X \otimes \mathbb{1}_E) W_n$, is Heisenberg $B \not\to A$ semicausal. Then, the map $\Phi : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$, defined by $\Phi(V) = V^\dagger (X \otimes \mathbb{1}_E) W$, is also Heisenberg $B \not\to A$ semicausal.

Proof. For $X_A \in \mathcal{B}(\mathcal{H}_A)$ and $\rho \in \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B)$, we have that $\rho\, V_m^\dagger (X_A \otimes \mathbb{1}_B \otimes \mathbb{1}_E) \in \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_E; \mathcal{H}_A \otimes \mathcal{H}_B)$, since the trace-class operators are an ideal in the bounded operators. Hence, by definition of the ultraweak topology,

$$\mathrm{tr}\left[\rho\, V_m^\dagger (X_A \otimes \mathbb{1}_B \otimes \mathbb{1}_E) W \right] = \lim_{n \to \infty} \mathrm{tr}\left[\rho\, V_m^\dagger (X_A \otimes \mathbb{1}_B \otimes \mathbb{1}_E) W_n \right] = \lim_{n \to \infty} \mathrm{tr}\left[\rho\, (\Phi_{m,n}^A(X_A) \otimes \mathbb{1}_B) \right].$$

Since $\mathrm{tr}\left[\rho\, \Phi_{m,n}^A(X_A) \otimes \mathbb{1}_B \right]$ converges as $n \to \infty$ for every $\rho \in \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B)$, the sequence $\left(\Phi_{m,n}^A(X_A) \otimes \mathbb{1}_B \right)_{n \in \mathbb{N}}$ converges ultraweakly.[30] We call the limit $\Phi_m^A(X_A) \otimes \mathbb{1}_B$. It is then easy to see that $\Phi_m^A(X_A)$, viewed as a map on $\mathcal{B}(\mathcal{H}_A)$, is linear and continuous. This tells us that the map $\Phi_m : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$, defined by $\Phi_m(X) = V_m^\dagger (X \otimes \mathbb{1}_E) W$, is semicausal for all $m \in \mathbb{N}$. Furthermore, we have that $\rho^\dagger\, W^\dagger (X_A^\dagger \otimes \mathbb{1}_B \otimes \mathbb{1}_E) \in \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_E; \mathcal{H}_A \otimes \mathcal{H}_B)$ for all $X_A \in \mathcal{B}(\mathcal{H}_A)$ and $\rho \in \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B)$, and thus,

$$\mathrm{tr}\Big[\rho\, V^\dagger (X_A\otimes \mathbb{1}_B\otimes \mathbb{1}_E)W\Big] = \overline{\mathrm{tr}\Big[\rho^\dagger W^\dagger (X_A^\dagger\otimes \mathbb{1}_B\otimes \mathbb{1}_E)V\Big]} = \lim_{m\to\infty}\overline{\mathrm{tr}\Big[\rho^\dagger W^\dagger (X_A^\dagger\otimes \mathbb{1}_B\otimes \mathbb{1}_E)V_m\Big]} = \lim_{m\to\infty}\mathrm{tr}\Big[\rho\, V_m^\dagger (X_A\otimes \mathbb{1}_B\otimes \mathbb{1}_E)W\Big]$$
$$= \lim_{m\to\infty}\mathrm{tr}\Big[\rho\,(\Phi_m^A(X_A)\otimes \mathbb{1}_E)\Big].$$

Repeating the argument above then shows that Φ is semicausal. □

As a final preparatory step, we observe that, given a semicausal Lindblad generator, we can use its CP part to define a family of semicausal CP-maps.

Lemma V.12. Let $L:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$ be defined by $L(X):=V^\dagger(X\otimes\mathbb{1}_E)V-K^\dagger X-XK$, with $V\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_B\otimes\mathcal{H}_E)$ and $K\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$. If L is Heisenberg $B\not\to A$ semicausal, then the map $S_{Y,Z}:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$, defined by

$$S_{Y,Z}(X)=\big[V(Z\otimes\mathbb{1}_B)-(Z\otimes\mathbb{1}_B\otimes\mathbb{1}_E)V\big]^\dagger\,(X\otimes\mathbb{1}_E)\,\big[V(Y\otimes\mathbb{1}_B)-(Y\otimes\mathbb{1}_B\otimes\mathbb{1}_E)V\big],$$

is Heisenberg $B\not\to A$ semicausal for every $Y,Z\in\mathcal{B}(\mathcal{H}_A)$.

Proof. For every $M\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$, we define the map $\Psi_M:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$ by

$$\Psi_M(X)=L(M^\dagger XM)-M^\dagger L(XM)-L(M^\dagger X)M+M^\dagger L(X)M$$
$$=\big[(M\otimes\mathbb{1}_E)V-VM\big]^\dagger(X\otimes\mathbb{1}_E)\big[(M\otimes\mathbb{1}_E)V-VM\big].$$

This map has already been used, for a different purpose, in Lindblad's original work [Ref. 19, Eq. (5.1)]. It follows from the semicausality of L that if we choose $M=M_A\otimes\mathbb{1}_B$ for some $M_A\in\mathcal{B}(\mathcal{H}_A)$, then Ψ_M is semicausal. Furthermore, a calculation shows that

$$\frac{1}{4}\sum_{k=0}^{3} i^k\Psi_{M+i^kN}(X)=\big[VN-(N\otimes\mathbb{1}_E)V\big]^\dagger\,(X\otimes\mathbb{1}_E)\,\big[VM-(M\otimes\mathbb{1}_E)V\big].$$

By choosing $N=Z\otimes\mathbb{1}_B$ and $M=Y\otimes\mathbb{1}_B$, it follows that $S_{Y,Z}$ is the linear combination of four semicausal maps and, hence, is itself semicausal. □

We now combine this lemma with an integration over the Haar measure to obtain the key lemma in our proof.

Lemma V.13. Let $L:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$ be defined by $L(X):=V^\dagger(X\otimes\mathbb{1}_E)V-K^\dagger X-XK$, with $V\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_B\otimes\mathcal{H}_E)$ and $K\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$. If L is Heisenberg $B\not\to A$ semicausal, then there exists $B\in\mathcal{B}(\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_E)$ such that the map $S:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$, defined by

$$S(X)=\big[V-\mathbb{1}_A\otimes B\big]^\dagger(X\otimes\mathbb{1}_E)\big[V-\mathbb{1}_A\otimes B\big],$$

is also Heisenberg $B\not\to A$ semicausal.

Furthermore, if $\mathcal{H}_A$ is finite-dimensional, then we can choose $B=\mathrm{tr}_A[V]/\dim(\mathcal{H}_A)$.

Proof. Let $\mathcal{H}_n$ and $\mathcal{H}_m$ be n and m dimensional subspaces of $\mathcal{H}_A$ with respective orthogonal projections $P_n\in\mathcal{B}(\mathcal{H}_A)$ and $P_m\in\mathcal{B}(\mathcal{H}_A)$. Since for every $U\in\mathcal{U}_P(\mathcal{H}_n)$ and $W\in\mathcal{U}_P(\mathcal{H}_m)$, the map $S_{U,W}$, defined in Lemma V.12, is semicausal and also the map $\overline{S}:\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)\to\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$, defined by

$$\overline{S}(X):=\int_{\mathcal{U}_P(\mathcal{H}_n)}\int_{\mathcal{U}_P(\mathcal{H}_m)}(U\otimes\mathbb{1}_B)S_{U,W}(X)(W^\dagger\otimes\mathbb{1}_B)\,dWdU,$$

is semicausal. Writing out the definition of $S_{U,W}$ yields

$$\overline{S}(X)=\left[V(P_n\otimes\mathbb{1}_B)-\int_{\mathcal{U}_P(\mathcal{H}_n)}(U\otimes\mathbb{1}_B\otimes\mathbb{1}_E)V(U^\dagger\otimes\mathbb{1}_E)dU\right]^\dagger(X\otimes\mathbb{1}_E)\left[V(P_m\otimes\mathbb{1}_B)-\int_{\mathcal{U}_P(\mathcal{H}_m)}(W\otimes\mathbb{1}_B\otimes\mathbb{1}_E)V(W^\dagger\otimes\mathbb{1}_B)dW\right]$$
$$=\left[V(P_n\otimes\mathbb{1}_B)-P_n\otimes\frac{1}{n}\mathrm{tr}_{P_n}[V]\right]^\dagger(X\otimes\mathbb{1}_E)\left[V(P_m\otimes\mathbb{1}_B)-P_m\otimes\frac{1}{m}\mathrm{tr}_{P_m}[V]\right],$$

where the last line was obtained by using Lemma V.9. If $\mathcal{H}_A$ is finite-dimensional, we can choose $\mathcal{H}_n=\mathcal{H}_m=\mathcal{H}_A$ so that $P_n=P_m=\mathbb{1}_A$ and obtain the desired result immediately. If $\mathcal{H}_A$ is separable infinite-dimensional and $\{|e_i\rangle\}_{i\in\mathbb{N}}$ is an orthonormal basis and $\mathcal{H}_k:=\mathrm{span}\{|e_1\rangle,|e_2\rangle,\ldots,|e_k\rangle\}$, then by Lemma V.9, the sequence $\left(P_k\otimes\frac{1}{k}\mathrm{tr}_{P_k}[V]\right)_{k\in\mathbb{N}}$ has an ultraweakly convergent subsequence with a limit $\mathbb{1}_A\otimes B$, where $B\in\mathcal{B}(\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_E)$. Furthermore, since $(P_k)_{k\in\mathbb{N}}$ converges ultraweakly to $\mathbb{1}_A$, we have that the sequence $\left(V(P_k\otimes\mathbb{1}_B)-P_k\otimes\frac{1}{k}\mathrm{tr}_{P_k}[V]\right)_{k\in\mathbb{N}}$ has a subsequence that converges ultraweakly to $V-\mathbb{1}_A\otimes B$. Hence, by passing to subsequences, we can apply Lemma V.11, which yields that S is semicausal. □

Remark V.14. The previous two lemmas are at the heart of our result. They illustrate a (to the best of our knowledge) novel technique that allows to characterize GKLS generators with a certain constraint if this constraint is well understood for completely positive maps. It seems useful to develop this method more generally, but this is beyond the scope of the present work.

With these tools at hand, we can now prove our main result.

Proof (Theorem V.6). A straightforward calculation shows that L, defined via (22a) and (22b), is semicausal. To prove the converse, note that by the Stinespring dilation theorem, there exist a separable Hilbert space $\tilde{\mathcal{H}}_E$ and $\tilde{V} \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B; \mathcal{H}_A \otimes \mathcal{H}_B \otimes \tilde{\mathcal{H}}_E)$ such that $\Phi(X) = \tilde{V}^\dagger(X \otimes \mathbb{1}_E)\tilde{V}$. It is well known [see, e.g., Ref. 31 (Theorems 2.1 and 2.2)] that if $\mathcal{H}_A$ and $\mathcal{H}_B$ are finite-dimensional with dimensions d_A and d_B, then $\tilde{\mathcal{H}}_E$ can be chosen such that $\dim(\tilde{\mathcal{H}}_E) \leq (d_A d_B)^2$. By Lemma V.13, there exists $\tilde{B} \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_B \otimes \tilde{\mathcal{H}}_E)$ such that the map $\Phi_0 \in \mathrm{CP}_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$, defined by $\Phi_0(X) = \left[\tilde{V} - \mathbb{1}_A \otimes \tilde{B}\right]^\dagger (X \otimes \mathbb{1}_E)\left[\tilde{V} - \mathbb{1}_A \otimes \tilde{B}\right]$, is semicausal. We define $V_{sc} = \tilde{V} - \mathbb{1} \otimes \tilde{B}$ and obtain

$$\Phi(X_A \otimes \mathbb{1}_B) = \Phi_0(X_A \otimes \mathbb{1}_B) + \kappa^\dagger(X_A \otimes \mathbb{1}_B) + (X_A \otimes \mathbb{1}_B)\kappa,$$

where $\kappa = (\mathbb{1}_A \otimes \tilde{B}^\dagger)V_{sc} + \frac{1}{2}\left(\mathbb{1}_A \otimes \tilde{B}^\dagger\tilde{B}\right)$. Since L and Φ_0 are semicausal, we can write $L(X_A \otimes \mathbb{1}) = L^A(X_A) \otimes \mathbb{1}_B$ and $\Phi_0(X_A \otimes \mathbb{1}_B) = \Phi_0^A(X_A) \otimes \mathbb{1}_B$ for all $X_A \in \mathcal{B}(\mathcal{H}_A)$. Hence,

$$L(X_A \otimes \mathbb{1}_B) - \Phi_0(X_A \otimes \mathbb{1}_B) = \left(L^A(X_A) - \Phi_0^A(X_A)\right) \otimes \mathbb{1}_B = -(K - \kappa)^\dagger(X_A \otimes \mathbb{1}_B) - (X_A \otimes \mathbb{1}_B)(K - \kappa). \tag{16}$$

It follows that the map defined by $X \mapsto -(K - \kappa)^\dagger X - X(K - \kappa)$ is semicausal. Thus, Lemma V.10 implies that there exist $K_A \in \mathcal{B}(\mathcal{H}_A)$ and a self-adjoint $H_B \in \mathcal{B}(\mathcal{H}_B)$ such that $K - \kappa = K_A \otimes \mathbb{1} + \mathbb{1} \otimes iH_B$.

What we have achieved so far is that $\tilde{V} = V_{sc} + \mathbb{1} \otimes \tilde{B}$ and $K = (\mathbb{1}_A \otimes \tilde{B}^\dagger)V_{sc} + \frac{1}{2}\mathbb{1} \otimes \tilde{B}^\dagger\tilde{B} + K_A \otimes \mathbb{1} + \mathbb{1} \otimes iH_B$. Hence, if we can decompose $V_{sc} = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$, then we are basically done. However, this decomposition is given (up to details) by the equivalence between semicausal and semilocalizable channels.[10] Since the conclusion in Ref. 10 was in the finite-dimensional setting, we will repeat the argument here, showing that it goes through also for infinite-dimensional spaces while paying special attention to the dimensions of the spaces involved. Since $\Phi_0 \in \mathrm{CP}_\sigma(\mathcal{H}_A \otimes \mathcal{H}_B)$ and $\Phi_0(X_A \otimes \mathbb{1}_B) = \Phi_0^A(X_A) \otimes \mathbb{1}_B$, we also have $\Phi_0^A \in \mathrm{CP}_\sigma(\mathcal{H}_A)$. By the Stinespring dilation theorem (for normal CP-maps), there exist a separable Hilbert space $\mathcal{H}_F$ and $W \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_F)$ such that $\Phi_0^A(X_A) = W^\dagger(X_A \otimes \mathbb{1}_F)W$ and such that $\mathrm{span}\{(X_A \otimes \mathbb{1}_F)W|\psi\rangle | X_A \in \mathcal{B}(\mathcal{H}_A), |\psi\rangle \in \mathcal{H}_A\}$ is dense in $\mathcal{H}_A \otimes \mathcal{H}_F$. The last condition is called the minimality condition. We then get

$$V_{sc}^\dagger(X_A \otimes \mathbb{1}_B \otimes \mathbb{1}_{\tilde{E}})V_{sc} = (W \otimes \mathbb{1}_B)^\dagger(X_A \otimes \mathbb{1}_F \otimes \mathbb{1}_B)(W \otimes \mathbb{1}_B).$$

Clearly, $\mathrm{span}\{(X_A \otimes \mathbb{1}_F \otimes \mathbb{1}_B)(W \otimes \mathbb{1}_B)|\psi\rangle | X_A \in \mathcal{B}(\mathcal{H}_A), |\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B\}$ is dense in $\mathcal{H}_A \otimes \mathcal{H}_F \otimes \mathcal{H}_B$. Thus, by minimality, there exists an isometry $\tilde{U} \in \mathcal{B}(\mathcal{H}_F \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \tilde{\mathcal{H}}_E)$ such that $V_{sc} = (\mathbb{1}_A \otimes \tilde{U})(W \otimes \mathbb{1}_B)$. In the finite-dimensional case, the fact that $\tilde{U}$ is an isometry then implies that $\dim(\mathcal{H}_F) \leq \dim(\tilde{\mathcal{H}}_E)$ such that we can think of $\mathcal{H}_F$ as a subspace of $\tilde{\mathcal{H}}_E$. Thus, $\tilde{U}$ can be extended to a unitary operator $\hat{\tilde{U}} \in \mathcal{U}(\tilde{\mathcal{H}}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \tilde{\mathcal{H}}_E)$. Then, defining $\mathcal{H}_E = \tilde{\mathcal{H}}_E$, $U = \hat{\tilde{U}}$, $B = \tilde{B}$, and $A = W$ proves the claim in this case. In the infinite-dimensional case, we can take $\mathcal{H}_E = \mathcal{H}_F \oplus \tilde{\mathcal{H}}_E$. We can now view both $\tilde{\mathcal{H}}_E \otimes \mathcal{H}_B$ and $\mathcal{H}_F \otimes \mathcal{H}_B$ as closed subspaces of $\mathcal{H}_E \otimes \mathcal{H}_B$. Then, $\left(\tilde{U}(\mathcal{H}_F \otimes \mathcal{H}_B)\right)^\perp$ and $(\mathcal{H}_F \otimes \mathcal{H}_B)^\perp$ are isomorphic. Hence, $\tilde{U}$ can be extended to a unitary operator $\hat{\tilde{U}} \in \mathcal{U}(\mathcal{H}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$. We finish the proof by defining $U = \hat{\tilde{U}}$, $B = (\mathbb{1}_B \otimes \mathbb{1}_{\tilde{E}\to E})\tilde{B}$, and $A = (\mathbb{1}_A \otimes \mathbb{1}_{F\to E})W$, where $\mathbb{1}_{\tilde{E}\to E}$ and $\mathbb{1}_{F\to E}$ denote the isometric embeddings of $\tilde{\mathcal{H}}_E$ and $\mathcal{H}_F$ into $\mathcal{H}_E$, respectively. □

As a first consequence, we obtain the analogous theorem for semigroups of Schrödinger $B \not\to A$ semicausal CP-maps.

Corollary V.15. Let $L : \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B)$ be defined by $L(\rho) = \Phi_S(\rho) - K\rho - \rho K^\dagger$, where $\Phi_S \in \mathrm{CP}_S(\mathcal{H}_A \otimes \mathcal{H}_B)$, with $\Phi_S(\rho) = \mathrm{tr}_E\left[V\rho V^\dagger\right]$ and $K \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$. Then, L is Schrödinger $B \not\to A$ semicausal if and only if K, V, and $\mathcal{H}_E$ can be chosen as in (22a) and (22b).

As a further corollary, we translate the results above to the familiar representation in terms of jump-operators (by going from Stinespring to Kraus).

Corollary V.16. A map $L : \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{S}_1(\mathcal{H}_A \otimes \mathcal{H}_B)$ generates a (trace-)norm-continuous semigroup of trace-preserving Schrödinger $B \not\to A$ semicausal CP-maps if and only if there exist $\{\phi_j\}_j \subset \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$, $\{B_j\}_j \subset \mathcal{B}(\mathcal{H}_B)$, $H_A \in \mathcal{B}(\mathcal{H}_A)$, and $H_B \in \mathcal{B}(\mathcal{H}_B)$ such that $\{\phi_j\}_j$ is a set of Kraus operators of a Schrödinger $B \not\to A$ semicausal CP-map and $\{B_j\}_j$ is a set of Kraus operators of some CP-map such that

$$L(\rho) = -i[H_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes H_B, \rho] + \sum_j (\phi_j + \mathbb{1}_A \otimes B_j)\rho(\phi_j + \mathbb{1}_A \otimes B_j)^\dagger - \frac{1}{2}\left\{\mathbb{1}_A \otimes B_j^\dagger B_j + \phi_j^\dagger\phi_j, \rho\right\} - (\mathbb{1}_A \otimes B_j^\dagger)\phi_j\rho - \rho\phi_j^\dagger(\mathbb{1}_A \otimes B_j).$$

Proof. A simple calculation by defining the Kraus operators as $(\mathbb{1}_{AB} \otimes |e_i\rangle)V$, with $\{|e_j\rangle\}_j$ being an orthonormal basis of $\mathcal{H}_E$ and V given by Theorem V.6. □

We conclude this section about semicausal semigroups with an example that uses our normal form in full generality.

Example. We consider the scenario of two 2-level atoms that can interact according to the processes specified in Fig. 6. We can describe this process either via a dilation (as in Theorem V.6) or via the Kraus operators (as in Corollary V.16). In the dilation picture, we introduce an auxiliary Hilbert space $\mathcal{H}_E := \mathcal{H}_1 \otimes \mathcal{H}_2$, where $\mathcal{H}_i$ is for the ith photon. Then, the process is described by $V = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B) + (\mathbb{1}_A \otimes B)$, with

$$\begin{aligned} A \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_E), \;\; A &= |0\rangle\langle 1|_A \otimes |11\rangle_E, \\ B \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E), \;\; B &= |10\rangle_E \otimes |0\rangle\langle 1|_B, \\ U \in \mathcal{U}(\mathcal{H}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E), \;\; U &= \mathbb{F}_{E;B}(\mathbb{1}_{\mathcal{H}_1} \otimes \tilde{U}), \end{aligned}$$

where $\tilde{U} \in \mathcal{U}(\mathcal{H}_2 \otimes \mathcal{H}_B)$ is determined by

$$\tilde{U}|00\rangle_{\mathcal{H}_2 B} = |00\rangle_{\mathcal{H}_2 B}, \;\; \tilde{U}|10\rangle_{\mathcal{H}_2 B} = |01\rangle_{\mathcal{H}_2 B}, \;\; \tilde{U}|11\rangle_{\mathcal{H}_2 B} = |11\rangle_{\mathcal{H}_2 B}.$$

The crucial feature of this example is that the CP-part of the generator $(\mathrm{tr}_E[V \cdot V^\dagger])$ cannot be written as a convex combination of the two building blocks (Φ_{sc} and $\mathrm{id}_A \otimes \hat{B}$). As mentioned also in the quantization procedure before, this is a pure quantum feature and stems from the fact that it cannot be determined if a photon arriving at the detector D_1 came from B or A. Hence, the system remains in a superposition state.

We can also look at the usual representation via jump operators. This can be achieved by switching from dilations to Kraus operators. We obtain the two jump-operators

$$L_1 := \underbrace{L_e \otimes L_a}_{=:\phi_1} + \mathbb{1}_A \otimes \underbrace{L_e}_{B_1}, \;\; L_2 := \underbrace{L_e \otimes |1\rangle\langle 1|}_{=:\phi_2},$$

where $L_e = |0\rangle\langle 1|$ and $L_a = L_e^\dagger$ describe emission and absorption of a photon, respectively. Thus, the usual Lindblad equation reads

$$\frac{d\rho}{dt} = (L_e \otimes L_a + \mathbb{1}_A \otimes L_e)\rho(L_e \otimes L_a + \mathbb{1}_A \otimes L_e) + (\mathbb{1}_A \otimes L_e)\rho(\mathbb{1}_A \otimes L_e) - \frac{1}{2}\left\{\mathbb{1}_A \otimes L_e^\dagger L_e + L_e^\dagger L_e \otimes \mathbb{1}_B, \rho\right\}.$$

It is also possible and instructive to consider the reduced dynamics on system A, which can also be described by a Lindblad equation, since B does not communicate to A (this is not true otherwise),

$$\frac{d\rho_A}{dt} = L_e \rho_A L_e^\dagger - \frac{1}{2}\left\{L_e^\dagger L_e, \rho_A\right\},$$

where $\rho_A(t) = \mathrm{tr}_B[\rho(t)]$. Not surprisingly (given our model), this describes an atom emitting photons.

C. Generators of semigroups of quantum superchannels

We finally turn to semigroups of quantum superchannels (on finite-dimensional spaces), that is, a collection of quantum superchannels $\{\hat{S}_t\}_{t\geq 0} \subseteq \mathcal{B}(\mathcal{B}(\mathcal{B}(\mathcal{H}_A); \mathcal{B}(\mathcal{H}_B)))$, such that $\hat{S}_0 = \mathrm{id}$, $\hat{S}_{t+s} = \hat{S}_t \hat{S}_s$, and the map $t \mapsto \hat{S}_t$ is continuous [with respect to any and, thus, all of the equivalent norms on the finite-dimensional space $\mathcal{B}(\mathcal{B}(\mathcal{B}(\mathcal{H}_A); \mathcal{B}(\mathcal{H}_B)))$]. To formulate a technically slightly stronger result, we call a map $\hat{S} \in \mathcal{B}(\mathcal{B}(\mathcal{B}(\mathcal{H}_A); \mathcal{B}(\mathcal{H}_B)))$ a preselecting supermap if $\mathfrak{C}_{A;B} \circ \hat{S} \circ \mathfrak{C}_{A;B}^{-1}$ is a Schrödinger $B \not\to A$ semicausal CP-map. Theorem V.3 then tells us that a superchannel is a special preselecting supermap. Again, as for semicausal CP-maps, we characterize the generators of semigroups of preselecting supermaps and superchannels in two ways: First, we answer how to determine if a given map $\hat{L} \in \mathcal{B}(\mathcal{B}(\mathcal{B}(\mathcal{H}_A); \mathcal{B}(\mathcal{H}_B)))$ is such a generator. Second, we provide a normal form for all generators.

The answer to the first question is really a corollary of Lemma V.5 together with Theorem V.3. To this end, define $\hat{\mathfrak{L}} := \mathfrak{C}_{AB;AB}\left(\mathfrak{C}_{A;B} \circ \hat{L} \circ \mathfrak{C}_{A;B}^{-1}\right) \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$, where we fix some orthonormal bases $\{|a_i\rangle\}_{i=1}^{\dim(\mathcal{H}_A)}$ and $\{|b_j\rangle\}_{j=1}^{\dim(\mathcal{H}_B)}$ of $\mathcal{H}_A$ and $\mathcal{H}_B$ such that $\mathfrak{C}_{A;B}$ is defined with respect to $\{|a_i\rangle\}_{i=1}^{\dim(\mathcal{H}_A)}$ and $\mathfrak{C}_{AB;AB}$ is defined with respect to the product of the two bases. Furthermore, we introduced the spaces $\mathcal{H}_{A_1} = \mathcal{H}_{A_2} = \mathcal{H}_A$ and $\mathcal{H}_{B_1} = \mathcal{H}_{B_2} = \mathcal{H}_A$ for notational convenience. Finally, we define $P^\perp \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$ to be the orthogonal projection onto the orthogonal complement of $\{|\Omega\rangle\}$, where $|\Omega\rangle = \sum_{i,j} |a_i\rangle \otimes |b_j\rangle \otimes |a_i\rangle \otimes |b_j\rangle$. We then have the following lemma:

Lemma V.17. A linear map $\hat{L} \in \mathcal{B}(\mathcal{B}(\mathcal{B}(\mathcal{H}_A); \mathcal{B}(\mathcal{H}_B)))$ generates a semigroup of quantum superchannels if and only if

- *$\hat{\mathfrak{L}}$ is self-adjoint and $P^\perp \hat{\mathfrak{L}} P^\perp \geq 0$,*

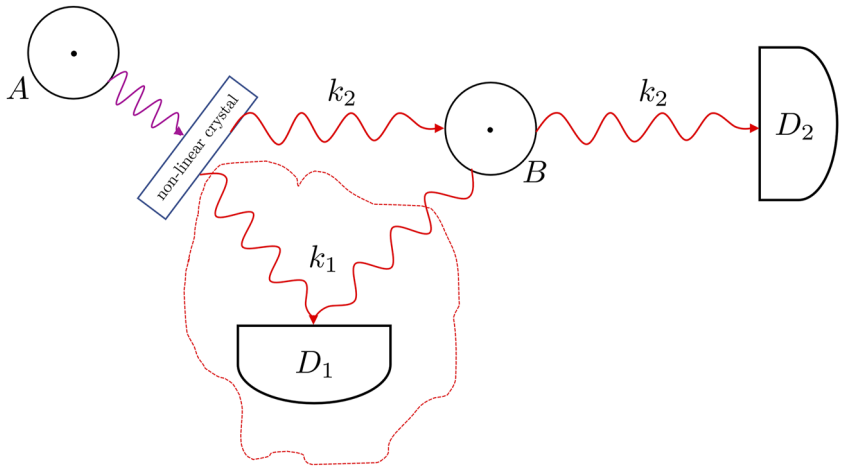


FIG. 6. Systems A and B describe 2-level systems, respectively. The allowed interactions are infinitesimally described as follows: If A is in its excited state, it can emit a photon. Through parametric down-conversion, the photon is converted into two photons (of lower energy). One of those two photons, k_1, is sent to a detector D_1. The other, k_2, is sent to B. If B is in its ground state, it absorbs k_2. If B is in its excited state, it cannot absorb k_2, so k_2 passes through B and travels to a detector D_2. Additionally, in this case, B can emit a photon, indistinguishable from k_1, to D_1.

- $(\mathbb{F}_{A_1;B_1} \otimes \mathbb{1}_{A_2})\mathrm{tr}_{B_2}\left[\hat{\mathfrak{L}}\right](\mathbb{F}_{A_1;B_1} \otimes \mathbb{1}_{A_2}) = \mathbb{1}_{B_1} \otimes \hat{\mathfrak{L}}^A$ *for some (then necessarily self-adjoint)* $\hat{\mathfrak{L}}^A \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{A_2})$, *and*
- $\mathrm{tr}_{A_1}\left[\hat{\mathfrak{L}}^A\right] = 0$. $\hat{L}$ *is preselecting if and only if the first two conditions hold.*

Proof. Theorem V.3 tells us that $\{\hat{S}_t\}_{t\geq 0}$ forming a semigroup of superchannels is eqiuvalent to $S_t = \mathfrak{C}_{A;B} \circ \hat{S}_t \circ \mathfrak{C}_{A;B}^{-1}$ forming a semigroup of Schrödinger $B \not\to A$ semicausal CP-maps and that the reduced map S_t^A satisfies $S_t^A(\mathbb{1}_A) = \mathbb{1}_A$. By Lemma V.5, the semicausal semigroup property is equivalent to the first two conditions in the statement. This proves the claim about preselecting $\hat{L}$.

By differentiation, it follows that $S_t^A(\mathbb{1}_A) = \mathbb{1}_A$ is satisfied if and only if L^A, the generator of $\{S_t^A\}_{t\geq 0}$, satisfies $L^A(\mathbb{1}_A) = 0$. However, since $\mathrm{tr}_{A_1}\left[\hat{\mathfrak{L}}^A\right] = L^A(\mathbb{1}_A)$, the claim follows. □

We finally turn to a normal form for generators of semigroups of preselecting supermaps and superchannels.

Theorem V.18. *A linear map* $\hat{L} : \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B)) \to \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ *generates a semigroup of hyper-preselecting supermaps if and only if there exist a Hilbert space* $\mathcal{H}_E$, *a state* $\sigma \in \mathcal{B}(\mathcal{H}_E)$, *a unitary* $U \in \mathcal{U}(\mathcal{H}_B \otimes \mathcal{H}_E)$, *a self-adjoint operator* $H_B \in \mathcal{B}(\mathcal{H}_B)$, *and arbitrary operators* $A \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_E)$, $B \in \mathcal{B}(\mathcal{H}_B \otimes \mathcal{H}_E)$, *and* $K_A \in \mathcal{B}(\mathcal{H}_A)$ *such that* $\hat{L}$ *acts on* $T \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ *as* $\hat{L}(T) = \hat{\Phi}(T) - \hat{\kappa}_L(T) - \hat{\kappa}_R(T)$ *with*

$$\begin{aligned}\hat{\Phi}(T)(\rho) = \mathrm{tr}_E\Big[U\ (T \otimes \mathrm{id}_E)(A(\rho\otimes\sigma)A^\dagger)\ U^\dagger\Big] + \mathrm{tr}_E\Big[B\ (T \otimes \mathrm{id}_E)((\rho\otimes\sigma)A^\dagger)\ U^\dagger\Big] \\ + \mathrm{tr}_E\Big[U\ (T \otimes \mathrm{id}_E)(A(\rho\otimes\sigma))\ B^\dagger\Big] + \mathrm{tr}_E\Big[B\ (T \otimes \mathrm{id}_E)((\rho\otimes\sigma))\ B^\dagger\Big],\end{aligned} \tag{17}$$

$$\hat{\kappa}_L(T)(\rho) = \mathrm{tr}_E\Big[B^\dagger U\ (T \otimes \mathrm{id}_E)(A(\rho\otimes\sigma))\Big] + \frac{1}{2}\mathrm{tr}_E\Big[B^\dagger B(T \otimes \mathrm{id}_E)(\rho\otimes\sigma)\Big]\ + T(K_A\,\rho) + iH_B\,T(\rho), \tag{18a}$$

$$\hat{\kappa}_R(T)(\rho) = \mathrm{tr}_E\Big[(T \otimes \mathrm{id}_E)((\rho\otimes\sigma)A^\dagger)\ U^\dagger B\Big] + \frac{1}{2}\mathrm{tr}_E\Big[(T \otimes \mathrm{id}_E)(\rho\otimes\sigma)B^\dagger B\Big]\ + T(\rho\,K_A^\dagger) - T(\rho)\,iH_B. \tag{18b}$$

We can choose σ *to be pure and* $\mathcal{H}_E$ *with* $\dim(\mathcal{H}_E) \leq (d_A d_B)^2$, *where* d_A *and* d_B *are the dimensions of* $\mathcal{H}_A$ *and* $\mathcal{H}_B$, *respectively. Furthermore,* $\hat{L}$ *generates a semigroup of superchannels if and only if* $\hat{L}$ *generates a semigroup of preselecting supermaps and* $\mathrm{tr}_\sigma\left[A^\dagger A\right] = K_A + K_A^\dagger$. *In that case, we can split* $\hat{L}$ *into a dissipative part* $\hat{D}$ *and a "Hamiltonian" part* $\hat{H}$, *i.e., a part that generates a (semi-)group of invertible superchannels whose inverses are superchannels as well. We have* $\hat{L}(T) = \hat{D}(T) + \hat{H}(T)$, *with*

$$\hat{D}(T)(\rho) = \mathrm{tr}_E\left[\hat{D}'(T)(\rho)\right] \quad \textit{and} \quad \hat{H}(T)(\rho) = -i[H_B\,,\,T(\rho)] - iT([H_A\,,\,\rho]),$$

where H_A *is the imaginary part of* K_A, *where*

$$\hat{D}'(T)(\rho) = U(T\otimes \mathrm{id}_E)(A(\rho\otimes\sigma)A^\dagger)U^\dagger - \frac{1}{2}(T\otimes \mathrm{id}_E)(\{A^\dagger A\,,\,\rho\otimes\sigma\}) \tag{19a}$$

$$+ B(T\otimes \mathrm{id}_E)(\rho\otimes\sigma)B^\dagger - \frac{1}{2}\{B^\dagger B\,,\,(T\otimes \mathrm{id}_E)(\rho\otimes\sigma)\} \tag{19b}$$

$$+ \left[U(T\otimes \mathrm{id}_E)(A(\rho\otimes\sigma))\,,\,B^\dagger\right] + \left[B\,,\,(T\otimes \mathrm{id}_E)((\rho\otimes\sigma)A^\dagger)U^\dagger\right] \tag{19c}$$

and where $[\cdot,\cdot]$ *and* $\{\cdot,\cdot\}$ *denote the commutator and anticommutator, respectively.*

Remark V.19. Similar to Theorem V.6, the Proof of Theorem V.18 is constructive. In Appendix D, we discuss in detail how to obtain the operators A, U, K_A, B, H_A, and H_B starting from the conditions in Theorem V.17.

As in the classical case, the proof strategy is to use the relation between superchannels and semicausal channels and Theorem V.6. As this translation process is more involved than in the classical case, we need two auxiliary lemmas.

Lemma V.20. Let $S : \mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$ *be given by*

$$S(X) = \mathrm{tr}_E\left[(\mathbb{1}_A\otimes L_B)(L_A\otimes\mathbb{1}_B)X(R_A^\dagger\otimes\mathbb{1}_B)(\mathbb{1}_A\otimes R_B^\dagger)\right], \tag{20}$$

with Hilbert spaces $\mathcal{H}_C$ *and* $\mathcal{H}_E$, *operators* $L_A, R_A \in \mathcal{B}(\mathcal{H}_A;\mathcal{H}_A\otimes\mathcal{H}_C)$, *and* $L_B, R_B \in \mathcal{B}(\mathcal{H}_C\otimes\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_E)$. *Then, for* $T \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A);\mathcal{B}(\mathcal{H}_B))$ *and* $\rho \in \mathcal{B}(\mathcal{H}_A)$,

$$\left[\mathfrak{C}_{A;B}^{-1}\circ S\circ\mathfrak{C}_{A;B}\right](T)(\rho) = \mathrm{tr}_E\left[V_L(T\otimes\mathrm{id}_C)\left(W_L\rho W_R^\dagger\right)V_R^\dagger\right], \tag{21}$$

with $V_L = L_B\mathbb{F}_{B;C}$, $V_R = R_B\mathbb{F}_{B;C}$, *and* $W_L = L_A^{T_A}$, $W_R = R_A^{T_A}$. *Here, the partial transpose on* $\mathcal{H}_A$ *is taken with respect to the basis used to define the Choi–Jamiołkowski isomorphism.*

Proof. The proof is a direct calculation. We present it in detail in Appendix A. □

Lemma V.21. Let $X \in \mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_C;\mathcal{H}_A\otimes\mathcal{H}_B)$, $Y \in \mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)$, $\rho \in \mathcal{S}_1(\mathcal{H}_B)$. *Then,* $\mathrm{tr}_\rho[XY]^T = \mathrm{tr}_C\left[Y^{T_A}(\mathbb{1}_A\otimes\rho)X^{T_A}\right]$.

Proof. The proof is a direct calculation. We present it in detail in Appendix A. □

We are finally ready to prove Theorem V.18

Proof (Theorem V.18). The idea is to relate the generators of superchannels to semicausal maps. This relation is given by definition for preselecting supermaps and by Theorem V.3 for superchannels. For a generator $\hat{L}$ of a semigroup of preselecting supermaps $\{\hat{S}_t\}_{t\geq 0}$, we have

$$\hat{L} = \mathfrak{C}_{A;B}^{-1}\circ\left.\frac{d}{dt}\right|_{t=0}[\mathfrak{C}_{A;B}\circ\hat{S}_t\circ\mathfrak{C}_{A;B}^{-1}]\circ\mathfrak{C}_{A;B}.$$

Thus, $\hat{L}$ generates a semigroup of preselecting supermaps if and only if $\hat{L}$ can be written as $\hat{L} = \mathfrak{C}_{A;B}^{-1}\circ L\circ\mathfrak{C}_{A;B}$ for some generator L of a semigroup of Schrödinger $B \not\to A$ semicausal CP-maps. Thus, to prove the first part of our theorem, we can take the normal form in Corollary V.15 and compute the similarity transformation above. We now execute this in detail. To start with, Corollary V.15 tells us that $L(\rho) = \Phi_S(\rho) - K\rho - \rho K^\dagger$, where

$$\Phi_S(\rho) = \mathrm{tr}_E\left[V\rho V^\dagger\right], \text{ with } V = (\mathbb{1}_A\otimes\tilde{U})(\tilde{A}\otimes\mathbb{1}_B) + (\mathbb{1}_A\otimes\tilde{B}), \tag{22a}$$

$$K = (\mathbb{1}_A\otimes\tilde{B}^\dagger\tilde{U})(\tilde{A}\otimes\mathbb{1}_B) + \frac{1}{2}\mathbb{1}_A\otimes\tilde{B}^\dagger\tilde{B} + \tilde{K}_A\otimes\mathbb{1}_B + \mathbb{1}_A\otimes i\tilde{H}_B, \tag{22b}$$

for some unitary $\tilde{U} \in \mathcal{U}(\mathcal{H}_E\otimes\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_E)$, some self-adjoint $\tilde{H}_B \in \mathcal{B}(\mathcal{H}_B)$, and some operators $\tilde{A} \in \mathcal{B}(\mathcal{H}_A;\mathcal{H}_A\otimes\mathcal{H}_E)$, $\tilde{B} \in \mathcal{B}(\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_E)$, and $\tilde{K}_A \in \mathcal{B}(\mathcal{H}_A)$. In order to apply Lemma V.20, we fix a unit vector $|\xi\rangle \in \mathcal{H}_E$ and define $\Xi_A := \mathbb{1}_A\otimes|\xi\rangle \in \mathcal{B}(\mathcal{H}_A;\mathcal{H}_A\otimes\mathcal{H}_E)$ and $\Xi_B := |\xi\rangle\otimes\mathbb{1}_B \in \mathcal{B}(\mathcal{H}_B;\mathcal{H}_E\otimes\mathcal{H}_B)$ so that $\mathbb{1}_A\otimes\tilde{B} = (\mathbb{1}_A\otimes\tilde{B}\Xi_B^\dagger)(\Xi_A\otimes\mathbb{1}_B)$. We can then write

$$\Phi_S(\rho) = \mathrm{tr}_E\left[(\mathbb{1}_A\otimes\tilde{U})(\tilde{A}\otimes\mathbb{1}_B)\rho(\tilde{A}^\dagger\otimes\mathbb{1}_B)(\mathbb{1}_A\otimes U^\dagger)\right] + \mathrm{tr}_E\left[(\mathbb{1}_A\otimes\tilde{B}\Xi_B^\dagger)(\Xi_A\otimes\mathbb{1}_B)\rho(\Xi_A^\dagger\otimes\mathbb{1}_B)(\mathbb{1}_A\otimes\Xi_B\tilde{B}^\dagger)\right]$$
$$+ \mathrm{tr}_E\left[(\mathbb{1}_A\otimes\tilde{U})(\tilde{A}\otimes\mathbb{1}_B)\rho(\Xi_A^\dagger\otimes\mathbb{1}_B)(\mathbb{1}_A\otimes\Xi_B\tilde{B}^\dagger)\right] + \mathrm{tr}_E\left[(\mathbb{1}_A\otimes\tilde{B}\Xi_B^\dagger)(\Xi_A\otimes\mathbb{1}_B)\rho(\tilde{A}^\dagger\otimes\mathbb{1}_B)(\mathbb{1}_A\otimes U^\dagger)\right],$$

which is an expression suitable for a term by term application of Lemma V.20. Doing so yields

$$\begin{aligned}\hat{\Phi}(T)(\rho) &:= (\mathfrak{C}_{A;B}^{-1} \circ \Phi_S \circ \mathfrak{C}_{A;B})(T)(\rho)\\ &= \mathrm{tr}_E\Big[U\ (T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma)A^\dagger)\ U^\dagger\Big] + \mathrm{tr}_E\Big[B\ (T \otimes \mathrm{id}_E)((\rho \otimes \sigma)A^\dagger)\ U^\dagger\Big]\\ &+ \mathrm{tr}_E\Big[U\ (T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma))\ B^\dagger\Big] + \mathrm{tr}_E\Big[B\ (T \otimes \mathrm{id}_E)((\rho \otimes \sigma))\ B^\dagger\Big],\end{aligned}$$

where we defined $U := \tilde{U}\mathbb{F}_{B;E}$, $B := \tilde{B}\Xi_B^\dagger\mathbb{F}_{B;E}$, $A := \tilde{A}^{T_A}\Xi_A^\dagger$, and $\sigma := |\xi\rangle\langle\xi|$. This proves Eq. (17). Similarly, upon defining $\kappa_L(\rho) := K\rho$, we can write[32]

$$\begin{aligned}\kappa_L(\rho) &= \mathrm{tr}_E\Big[(\mathbb{1}_A \otimes \mathbb{F}_{E;B}\Xi_B\tilde{B}^\dagger\tilde{U})(\tilde{A} \otimes \mathbb{1}_B)\rho(\Xi_A^\dagger \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes \mathbb{F}_{B;E})\Big] + \mathrm{tr}_E\Big[(\mathbb{1}_A \otimes \mathbb{F}_{E;B}\Xi_B\tilde{B}^\dagger\tilde{B}\Xi_B^\dagger)(\Xi_A \otimes \mathbb{1}_B)\rho(\Xi_A^\dagger \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes \mathbb{F}_{B;E})\Big]\\ &+ \mathrm{tr}_{\mathbb{C}}\Big[(\mathbb{1}_A \otimes \mathbb{1}_B)(\tilde{K}_A \otimes \mathbb{1}_B)\rho(\mathbb{1}_A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes \mathbb{1}_B)\Big] + \mathrm{tr}_{\mathbb{C}}[(\mathbb{1}_A \otimes iH_B)(\mathbb{1}_A \otimes \mathbb{1}_B)\rho(\mathbb{1}_A \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes \mathbb{1}_B)]\end{aligned}$$

and apply Lemma V.20 term by term, which yields

$$\begin{aligned}\hat{\kappa}_L(T)(\rho) &:= (\mathfrak{C}_{A;B}^{-1} \circ \kappa_L \circ \mathfrak{C}_{A;B})(T)(\rho)\\ &= \mathrm{tr}_E\Big[B^\dagger U\ (T \otimes \mathrm{id}_E)(A(\rho \otimes \sigma))\Big] + \frac{1}{2}\mathrm{tr}_E\Big[B^\dagger B(T \otimes \mathrm{id}_E)(\rho \otimes \sigma)\Big]\ + T(K_A\,\rho) + iH_B\,T(\rho),\end{aligned}$$

where U, A, and B are defined as above and $K_A := (\tilde{K}_A)^T$ and $H_B := \tilde{H}_B$. An analogous calculation with $\kappa_R(\rho) := \rho K^\dagger$ and $\hat{\kappa}_R(T) := (\mathfrak{C}_{A;B}^{-1} \circ \kappa_R \circ \mathfrak{C}_{A;B})(T)$ finishes the proof of the first part, since the claim about the dimension of $\mathcal{H}_E$ follows form the corresponding statements in Theorem V.6.

To prove the second part, first remember that we have observed above that Theorem V.3 implies that L is Schrödinger $B \not\to A$ semicausal, with $\mathrm{tr}_B[L(\rho)] = L^A(\mathrm{tr}_B[\rho])$. Furthermore, if we write $S_t = \mathfrak{C}_{A;B} \circ \hat{S}_t \circ \mathfrak{C}_{A;B}^{-1}$, then Theorem V.3 implies that S_t is Schrödinger $B \not\to A$ semicausal for all $t \geq 0$, with $\mathrm{tr}_B[S_t(\rho)] = S_t^A(\mathrm{tr}_B[\rho])$, and also $S_t^A(\mathbb{1}_A) = \mathbb{1}_A$ holds. Differentiating that expression at $t = 0$ yields the equivalent condition $L^A(\mathbb{1}_A) = 0$. Hence, our goal is to incorporate the last condition into the form of (22). To do so, we determine L^A by calculating $\mathrm{tr}_B[L(\rho)]$, where L is in the form of (22). We obtain $\mathrm{tr}_B[L(\rho)] = \mathrm{tr}_E\big[\tilde{A}\,\mathrm{tr}_B[\rho]\,\tilde{A}^\dagger\big] - \tilde{K}_A\mathrm{tr}_B[\rho] - \mathrm{tr}_B[\rho]\tilde{K}_A^\dagger$. Thus, the condition $L^A(\mathbb{1}) = 0$ holds if and only if $\mathrm{tr}_E\big[\tilde{A}\tilde{A}^\dagger\big] = \tilde{K}_A + \tilde{K}_A^\dagger$. Transposing both sides of this equation and using that the definition of A implies that $\tilde{A} = A^{T_A}\Xi_A$ yield $\big(\mathrm{tr}_E\big[A^{T_A}(\mathbb{1}_A \otimes \sigma)(A^\dagger)^{T_A}\big]\big)^T = K_A + K_A^\dagger$. However, the left-hand side is, by Lemma V.21, equal to $\mathrm{tr}_\sigma\big[A^\dagger A\big]$. This proves the claim that $\hat{L}$ generates a semigroup of superchannels if and only if $\hat{L}$ is hyper-preselecting and $\mathrm{tr}_\sigma\big[A^\dagger A\big] = K_A + K_A^\dagger$. Finally, defining $H_A := \frac{1}{2i}(K_A - K_A^\dagger)$ and a few rearrangements lead to (19). □

VI. CONCLUSION

A. Summary

The underlying question of this work is as follows: How can we mathematically characterize the processes that describe the aging of quantum devices? We have argued that, under a Markovianity assumption, such processes can be modeled by continuous semigroups of quantum superchannels. Therefore, the goal of this work was to provide a full characterization of such semigroups of superchannels.

We have derived such a general characterization in terms of the generators of these semigroups. Crucially, we have exploited that superchannels correspond to certain semicausal maps and that, therefore, it suffices to characterize generators of semigroups of semicausal maps. We have demonstrated both an efficient procedure for checking whether a given generator is indeed a valid semicausal GKLS generator and a complete characterization of such valid semicausal GKLS generators. The latter is constructive in the sense that it can be used to describe parametrizations of these generators. Aside from the theoretical relevance of these results, they will be valuable in studying properties of these generators numerically. Finally, we have translated these results back to the level of superchannels, thus answering our initial question.

We have also posed and answered the classical counterpart of the above question. That is, we have characterized the generators semigroups of classical superchannels and of semicausal non-negative maps. These results for the classical case might be of independent interest. From the perspective of quantum information theory, they provide a comparison helpful to understand and interpret the characterizations in the quantum case.

B. Outlook and open questions

We conclude by presenting some open questions raised by our work. First, in our proof of the characterization of semicausal GKLS generators, we have described a procedure for constructing a semicausal CP-map associated with such a generator. We believe that this method can be applied to a wide range of problems. Determining the exact scope of this method is currently work in progress.

Second, there is a wealth of results on the spectral properties of quantum channels and, in particular, semigroups of quantum channels. With the explicit form of generators of semigroups of superchannels now known, we can conduct analogous studies for semigroups of quantum superchannels. Understanding such spectral properties, and potentially how they differ from the properties in the scenario of quantum

channels, would, in particular, lead to a better understanding of the asymptotic behavior of semigroups of superchannels, e.g., with respect to entropy production,[33,34] the thermodynamics of quantum channels,[35] or entanglement-breaking properties.[36]

A further natural question would be a quantum superchannel analog of the Markovianity problem: When can a quantum superchannel $\hat{S}$ be written as $e^{\hat{L}}$ for some $\hat{L}$ that generates a semigroup of superchannels? Several works have investigated the Markovianity problem for quantum channels[21,37–39] and a divisibility variant of this question, both for quantum channels and for stochastic matrices.[40–42] It would be interesting to see how these results translate to quantum or classical superchannels. Similarly, we can now ask questions of reachability along Markovian paths. Yet another question aiming at understanding Markovianity is as follows: If we consider master equations arising from a Markovianity assumption on the underlying process formalized not via semigroups of channels but instead via semigroups of superchannels, what are the associated classes of (time-dependent) generators and corresponding CPTP evolutions?

Two related directions, both of which will lead to a better understanding of Markovian structures in higher order quantum operations, are as follows: support our mathematical characterization of the generators of semigroups of superchannels by a physical interpretation, similar to the Monte Carlo wave function interpretation of Lindblad generators of quantum channels, and extend our characterization from superchannels to general higher order maps.

This work has focused on generators of general semigroups of superchannels, without further restrictions. For quantum channels and their Lindblad generators, there exists a well-developed theory of locality, at the center of which are Lieb–Robinson bounds.[43] If we put locality restrictions on generators of superchannels, how do these translate to the generated superchannels?

Finally, an important conceptual direction for future work is to identify further applications of our theory of dynamical semigroups of superchannels. In the Introduction, we gave a physical meaning to semigroups of superchannels by relating them to the decay process of quantum devices. This, however, is only one possible interpretation. For example, semigroups of superchannels might also describe a manufacturing process, where a quantum device is created layer-by-layer. We hope that other use-cases will be found in the future.

ACKNOWLEDGMENTS

M.C.C. and M.H. thank Michael M. Wolf for insightful discussions about the contents of this paper. We also thank Li Gao, Lisa Hänggli, Robert König, and Farzin Salek for helpful suggestions for improving the presentation. M.C.C. and M.H. also thank the anonymous reviewers from TQC 2022 and from the Journal of Mathematical Physics for their constructive criticism. M.H. was supported by the Bavarian excellence network ENB via the International Ph.D. Program of Excellence *Exploring Quantum Matter* (EXQM). M.C.C. gratefully acknowledges support from the TopMath Graduate Center of the TUM Graduate School at the Technische Universität München, Germany, from the TopMath Program at the Elite Network of Bavaria, and from the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

APPENDIX A: PROOF OF LEMMAS V.20 AND V.21

In this appendix, we provide a complete proof of Lemmas V.20 and V.21.

Lemma A.1 (restatement of Lemma V.20). Let $S : \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$ *be given by*

$$S(X) = \mathrm{tr}_E\Big[(\mathbb{1}_A \otimes L_B)(L_A \otimes \mathbb{1}_B)X(R_A^\dagger \otimes \mathbb{1}_B)(\mathbb{1}_A \otimes R_B^\dagger)\Big],$$

with Hilbert spaces $\mathcal{H}_C$ *and* $\mathcal{H}_E$, *operators* $L_A, R_A \in \mathcal{B}(\mathcal{H}_A; \mathcal{H}_A \otimes \mathcal{H}_C)$, *and* $L_B, R_B \in \mathcal{B}(\mathcal{H}_C \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$. *Then, for* $T \in \mathcal{B}(\mathcal{B}(\mathcal{H}_A); \mathcal{B}(\mathcal{H}_B))$ *and* $\rho \in \mathcal{B}(\mathcal{H}_A)$,

$$\big[\mathfrak{C}_{A;B}^{-1} \circ S \circ \mathfrak{C}_{A;B}\big](T)(\rho) = \mathrm{tr}_E\Big[V_L(T \otimes \mathrm{id}_C)\Big(W_L \rho W_R^\dagger\Big)V_R^\dagger\Big],$$

with $V_L = L_B \mathbb{F}_{B;C}$, $V_R = R_B \mathbb{F}_{B;C}$ *and* $W_L = L_A^{T_A}$, $W_R = R_A^{T_A}$. *Here, the partial transpose on* $\mathcal{H}_A$ *is taken with respect to the basis used to define the Choi–Jamiołkowski isomorphism.*

Proof. Let $\{|e_i\rangle\}_i$ be the orthonormal basis of $\mathcal{H}_A$ with respet to which the Choi–Jamiołkowski isomorphism is defined. Let $\{|c_n\rangle\}_n$ be an orthonormal basis of $\mathcal{H}_C$. Then, the formal calculation, which is an algebraic version of drawing the corresponding tensor-network pictures,

can be executed as follows:

$$\begin{aligned}
\left[\mathfrak{C}_{A;B}^{-1}\circ S\circ\mathfrak{C}_{A;B}\right](T)(\rho) &= \mathrm{tr}_A\Big[(\rho^T\otimes\mathbb{1}_B)\,\mathrm{tr}_E\Big[(\mathbb{1}_A\otimes L_B)(L_A\otimes\mathbb{1}_B)\mathfrak{C}_{A;B}(T)(R_A^\dagger\otimes\mathbb{1}_B)(\mathbb{1}_A\otimes R_B^\dagger)\Big]\Big]\\
&= \mathrm{tr}_E\Big[L_B\,\mathrm{tr}_A\Big[(\rho^T\otimes\mathbb{1}_C\otimes\mathbb{1}_B)(L_A\otimes\mathbb{1}_B)\mathfrak{C}_{A;B}(T)(R_A^\dagger\otimes\mathbb{1}_B)\Big]R_B^\dagger\Big]\\
&= \sum_{i,j}\mathrm{tr}_E\Big[L_B\left(\mathrm{tr}_A\Big[(\rho^T\otimes\mathbb{1}_C)L_A|e_i\rangle\langle e_j|R_A^\dagger\Big]\otimes T(|e_i\rangle\langle e_j|)\right)R_B^\dagger\Big]\\
&= \sum_{i,j,k,m,n}\langle e_k\,c_n|\Big((\rho^T\otimes\mathbb{1}_C)L_A|e_i\rangle\langle e_j|R_A^\dagger\Big)e_k\,c_m\rangle\;\mathrm{tr}_E\Big[L_B\left(|c_n\rangle\langle c_m|\otimes T(|e_i\rangle\langle e_j|)\right)R_B^\dagger\Big]\\
&= \sum_{i,j,m,n}\langle e_i|\Big(L_A^T(\rho\otimes|c_n\rangle\langle c_m|)\overline{R}_A\Big)e_j\rangle\;\mathrm{tr}_E\Big[L_B\left(|c_n\rangle\langle c_m|\otimes T(|e_i\rangle\langle e_j|)\right)R_B^\dagger\Big]\\
&= \sum_{m,n}\mathrm{tr}_E\Big[L_B\left(|c_n\rangle\langle c_m|\otimes T\Big(L_A^T(\rho\otimes|c_n\rangle\langle c_m|)\overline{R}_A\Big)\right)R_B^\dagger\Big]\\
&= \mathrm{tr}_E\left[L_B\mathbb{F}_{B;C}(T\otimes\mathrm{id}_C)\left(\left[\sum_n(\mathbb{1}_A\otimes|c_n\rangle)L_A^T(\mathbb{1}_A\otimes|c_n\rangle)\right]\rho\left[\sum_m(\mathbb{1}_A\otimes|c_m\rangle)R_A^T(\mathbb{1}_A\otimes|c_m\rangle)\right]^\dagger\right)\mathbb{F}_{B;C}R_B^\dagger\right]\\
&= \mathrm{tr}_E\Big[V_L(T\otimes\mathrm{id}_C)\Big(W_L\rho W_R^\dagger\Big)V_R^\dagger\Big].
\end{aligned}$$

□

Lemma A.2. Let $X\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_C;\mathcal{H}_A\otimes\mathcal{H}_B)$, $Y\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B;\mathcal{H}_A\otimes\mathcal{H}_C)$, $\rho\in\mathcal{S}_1(\mathcal{H}_B)$. *Then,* $\mathrm{tr}_\rho[XY]^T=\mathrm{tr}_C\big[Y^{T_A}(\mathbb{1}_A\otimes\rho)X^{T_A}\big]$.

Proof. Let $\{|a_i\rangle\}_i$ be the orthonormal basis with respect to which the transposition is taken. Using the general identity $\mathrm{tr}\big[M^T\big]=\mathrm{tr}[M]$, the definition of the trace with respect to a trace-class operator, and the cyclicity of the trace, we obtain, for every $\sigma\in\mathcal{S}_1(\mathcal{H}_A)$,

$$\begin{aligned}
\mathrm{tr}\Big[\sigma\mathrm{tr}_\rho[XY]^T\Big] &= \mathrm{tr}\Big[\sigma^T\mathrm{tr}_\rho[XY]\Big]\\
&= \mathrm{tr}\Big[(\sigma^T\otimes\rho)XY\Big]\\
&= \sum_{i,j,k}\mathrm{tr}\Big[(\langle a_i|\otimes\mathbb{1}_B)(\sigma^T\otimes\rho)(|a_j\rangle\langle a_j|\otimes\mathbb{1}_B)X(|a_k\rangle\langle a_k|\otimes\mathbb{1}_C)Y(|a_i\rangle\otimes\mathbb{1}_B)\Big]\\
&= \sum_{i,j,k}\mathrm{tr}\Big[(\langle a_j|\otimes\mathbb{1}_B)(\sigma\otimes\rho)(|a_i\rangle\langle a_k|\otimes\mathbb{1}_B)X^{T_A}(|a_j\rangle\langle a_i|\otimes\mathbb{1}_C)Y^{T_A}(|a_k\rangle\otimes\mathbb{1}_B)\Big]\\
&= \sum_k\mathrm{tr}\left[\rho(\langle a_k|\otimes\mathbb{1}_B)X^{T_A}\left(\left(\sum_{i,j}\langle a_j|\sigma\,a_i\rangle|a_i\rangle\langle a_j|\right)\otimes\mathbb{1}_C\right)Y^{T_A}(|a_k\rangle\otimes\mathbb{1}_B)\right]\\
&= \mathrm{tr}\Big[(\mathbb{1}_A\otimes\rho)X^{T_A}(\sigma\otimes\mathbb{1}_C)Y^{T_A}\Big]\\
&= \mathrm{tr}\Big[\sigma\mathrm{tr}_C\Big[Y^{T_A}(\mathbb{1}_A\otimes\rho)X^{T_A}\Big]\Big].
\end{aligned}$$

This proves the claim. □

APPENDIX B: NO INFORMATION WITHOUT DISTURBANCE

Here, we prove a "no information without disturbance"-like lemma that yielded a useful interpretation in the main text.

Lemma B.1. Let $T\in\mathrm{CP}_\sigma(\mathcal{H}_A\otimes\mathcal{H}_B)$ *be such that*

$$T(X_A\otimes\mathbb{1}_B)=X_A\otimes\mathbb{1}_B \tag{B1}$$

for all $X_A\in\mathcal{B}(\mathcal{H}_A)$. *Then,* $T(X)=(\mathbb{1}_A\otimes W^\dagger)(X\otimes\mathbb{1}_E)(\mathbb{1}_A\otimes W)$ *for all* $X\in\mathcal{B}(\mathcal{H}_A\otimes\mathcal{H}_B)$ *and some isometry* $W\in\mathcal{B}(\mathcal{H}_B;\mathcal{H}_B\otimes\mathcal{H}_E)$, *where* $\mathcal{H}_E$ *is some Hilbert space.*

Proof. This claim follows from the uniqueness of the minimal Stinespring dilation in the same way as the "semicausal = semilocalizable" theorem. Write Eq. (B1) in the Stinespring form as

$$V^\dagger(X_A\otimes\mathbb{1}_B\otimes\mathbb{1}_E)V=X_A\otimes\mathbb{1}_B$$

for some $V \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B; \mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C)$. Then, V and $\mathbb{1}_{AB}$ are the Stinespring operators of the same CP-map $(X_A \mapsto X_A \otimes \mathbb{1}_B)$ and the latter clearly belongs to a minimal dilation. Thus, there exists an isometry $W \in \mathcal{B}(\mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$ such that $V = (\mathbb{1}_A \otimes W)\mathbb{1}_{AB}$. This is the claim. □

Note that the lemma above is just a formulation of the "obvious" fact that if system A undergoes a closed system evolution (id_A), then there is no interaction with an external system B.

APPENDIX C: CONSTRUCTIVE APPROACH TO THEOREM V.6

In this appendix, we are going to describe in detail how one can computationally construct the operators A, U, B, K_A, and H_B in Theorem V.6 if the conditions of Lemma V.5 are met.

Since it is important for an actual implementation on a computer, let us be very precise about notation. We introduce indexed copies of $\mathcal{H}_A$ and $\mathcal{H}_B$, i.e., $\mathcal{H}_{A_0} = \mathcal{H}_{A_1} = \mathcal{H}_{A_2} = \mathcal{H}_A$ and $\mathcal{H}_{B_0} = \mathcal{H}_{B_1} = \mathcal{H}_{B_2} = \mathcal{H}_B$. Furthermore, we fix orthonormal bases $\{|a_i\rangle\}_{i=1}^{d_A}$ and $\{|b_i\rangle\}_{i=1}^{d_B}$ of $\mathcal{H}_A$ and $\mathcal{H}_B$, respectively. We use the symbol Ω with some subscript to denote the maximally entangled state on various systems. For example, $|\Omega_{A_1;A_2}\rangle := \sum_i |a_i\rangle \otimes |a_i\rangle \in \mathcal{H}_{A_1} \otimes \mathcal{H}_{A_2}$ and $|\Omega_{A_1B_1;A_2B_2}\rangle = \sum_{i,j} |a_i\rangle \otimes |b_j\rangle \otimes |a_i\rangle \otimes |b_j\rangle \in \mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2}$. We further reserve $P \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$ for the orthogonal projection onto $\mathrm{span}\{|\Omega_{A_1B_1;A_2B_2}\rangle\}$ (i.e., $P = (d_A d_B)^{-1} |\Omega_{A_1B_1;A_2B_2}\rangle\langle\Omega_{A_1B_1;A_2B_2}|$) and take $P^\perp = \mathbb{1}_{A_1B_1A_2B_2} - P$.

Now, let $\mathfrak{L} \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$ be given as in Lemma V.5, then we can compute the operators A, U, B, K_A, and H_B via the following 15 steps:

1. Compute $\tau = P^\perp \mathfrak{L} P^\perp$.
2. Compute $V = (\mathbb{1}_{A_0B_0} \otimes \sqrt{\tau})(|\Omega_{A_0B_0;A_1B_1}\rangle \otimes \mathbb{1}_{A_2B_2})$.
3. Define $\mathcal{H}_E := \mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2}$ so that $V \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B; \mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_E)$ (identification).
4. Compute $B = \frac{1}{d_A} \mathrm{tr}_A[V]$.
5. Compute $V_{sc} = V - \mathbb{1}_A \otimes B$.
6. Compute $\tau_{sc} = (\mathbb{1}_{A_1B_1} \otimes V_{sc})^\dagger (|\Omega_{A_1B_1;AB}\rangle\langle\Omega_{A_1B_1;AB}| \otimes \mathbb{1}_E)(\mathbb{1}_{A_1B_1} \otimes V_{sc}) \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_A \otimes \mathcal{H}_B)$.
7. Choose any unit vector $|\beta\rangle \in \mathcal{H}_B$.
8. Compute $\tau_{sc}^A = (\mathbb{1}_{A_1A_2} \otimes \langle\beta|)\mathrm{tr}_{B_1}[\tau_{sc}](\mathbb{1}_{A_1A_2} \otimes |\beta\rangle)$.
9. Compute $\mathcal{H}_F = \mathrm{range}(\sqrt{\tau_{sc}^A})$ so that $\sqrt{\tau_{sc}^A} \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{A_2}; \mathcal{H}_F)$ is surjective.
10. Compute $A = (\mathbb{1}_{A_0} \otimes \sqrt{\tau_{sc}^A})(|\Omega_{A_0;A_1}\rangle \otimes \mathbb{1}_{A_2})$.
11. Compute U as *the* solution of the system of linear equations $\mathcal{M}(U) = V_{sc}$, where the $d_A^2 d_B^2 d_E \times d_F d_B^2 d_E$-matrix $\mathcal{M} : \mathcal{B}(\mathcal{H}_F \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E) \to \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B; \mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_E)$ is defined by $\mathcal{M}(U) = (\mathbb{1}_A \otimes U)(A \otimes \mathbb{1}_B)$. Clearly, we must first represent $\mathcal{M}$ with respect to some basis.
12. Compute $K = -\mathrm{tr}_{A_1B_1}\left[P\mathfrak{L}P^\perp + \frac{1}{2}\mathrm{tr}[P\mathfrak{L}]P\right]$, where we identify $\mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2} = \mathcal{H}_A \otimes \mathcal{H}_B$ so that $K \in \mathcal{B}(\mathcal{H}_A \otimes \mathcal{H}_B)$.
13. Compute $K_{sc} = K - (\mathbb{1}_A \otimes B^\dagger)V_{sc} - \frac{1}{2}\mathbb{1}_A \otimes B^\dagger B$.
14. Compute $K_A = \frac{1}{d_B}\mathrm{tr}_B[K_{sc}]$.
15. Compute $H_B = \frac{-i}{d_A}\mathrm{tr}_A[K_{sc} - K_A \otimes \mathbb{1}_B]$.

Note that the procedure above computes an isometry $U \in \mathcal{B}(\mathcal{H}_F \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$, which can then be extended to a unitary, if necessary. In that case, we also have to embed $\mathcal{H}_F$ into $\mathcal{H}_E$ and redefine A accordingly. More precisely, we need to execute the following additional steps:

16. Compute $\mathbb{1}_{F\to E} = \mathbb{1}_{A_1} \otimes |\beta\rangle_{B_1} \otimes \mathbb{1}_{A_2} \otimes |\beta\rangle_{B_2}$.
17. Redefine $A \leftarrow (\mathbb{1}_{A_0} \otimes \mathbb{1}_{F\to E})A$.
18. Extend U via the following steps:
    (a) Compute $\hat{U} = U(\mathbb{1}_{F\to E}^\dagger \otimes \mathbb{1}_B)$.
    (b) Compute an orthonormal basis $\{|f_i^\perp\rangle\}_{i=1}^N$ of $\mathrm{range}(\mathbb{1}_{EB} - \hat{U}^\dagger\hat{U})$.
    (c) Compute an orthonormal basis $\{|r_i^\perp\rangle\}_{i=1}^N$ of $\mathrm{range}(\mathbb{1}_{BE} - \hat{U}\hat{U}^\dagger)$.
    (d) Redefine $U \leftarrow \hat{U} + \sum_{i=1}^N |r_i^\perp\rangle\langle f_i^\perp|$.

Let us comment on why the steps above give the right result. In general, we have

$$\mathfrak{L} = P^\perp\mathfrak{L}P^\perp + P\mathfrak{L}P^\perp + P^\perp\mathfrak{L}P + P\mathfrak{L}P = \tau + \left(P\mathfrak{L}P^\perp + \frac{1}{2}\mathrm{tr}[P\mathfrak{L}]P\right) + \left(P^\perp\mathfrak{L}P + \frac{1}{2}\mathrm{tr}[P\mathfrak{L}]P\right).$$

Thus, the maps Φ and K appearing in the GKLS-form in Theorem V.6 can be extracted from the previous equation by applying the inverse of the Choi–Jamiołkowski isomorphism. One readily obtains $\Phi = \mathfrak{C}_{AB;AB}^{-1} \circ \tau$ and $K = -\mathrm{tr}_{A_1B_1}\left[P\mathfrak{L}P^\perp + \frac{1}{2}\mathrm{tr}[P\mathfrak{L}]P\right]$.

- Step 2 computes the Stinespring dilation of a CP-map whose Choi–Jamiołkowski operator is τ. A direct computation shows that $\tau = (\mathbb{1}_{A_1B_1} \otimes V)^\dagger (|\Omega_{A_1B_1;A_2B_2}\rangle\langle\Omega_{A_1B_1;A_2B_2}| \otimes \mathbb{1}_E)(\mathbb{1}_{A_1B_1} \otimes V)$.
- Step 4 computes the operator B in the representation. In the Proof of Theorem V.6, B was obtained from $\tilde{B}$, which, in turn, was obtained from V and Lemma V.13. In the finite-dimensional setting, Lemma V.13 constructs B exactly as is written down above.
- Steps 6, 7, and 8 define τ_{sc} as the Choi–Jamiołkowski operator of a CP-map with the Stinespring operator V_{sc}. Thus, according to the Proof of Theorem V.6, τ is the Choi–Jamiołkowski of a Heisenberg $B \not\to A$ semicausal map. Semicausality is expressed on the level of Choi–Jamiołkowski operators by the existence of an operator τ_{sc}^A such that $\mathrm{tr}_{B_1}[\tau_{sc}] = \tau_{sc}^A \otimes \mathbb{1}_{B_2}$ (compare with the Proof of Lemma V.5). Using this relation makes clear that step 8 extracts τ_{sc}^A from τ_{sc} and that the result is independent of the choice of $|\beta\rangle$.
- Step 10 defines A as the Stinespring dilation of the (reduced) map whose Choi–Jamiołkowski operator is τ_{sc}^A. The dilation constructed in this way is minimal. This is exactly the way in which the operator $W = A$ was constructed in the Proof of Theorem V.6.
- Step 11 obtains U by solving the defining relation (for $\tilde{U}$) in the Proof of Theorem V.6. One might wonder why the solution to this system of equations is unique (even though $\mathcal{M}$ is not a square matrix). Uniqueness follows from the minimality of $A \otimes \mathbb{1}_B$, that is, vectors of the form $(X_A \otimes \mathbb{1}_{FB})(A \otimes \mathbb{1}_B)|\psi\rangle$ span $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_E$. In detail, if U and U' satisfy $\mathcal{M}(U) = \mathcal{M}(U')$, then $0 = (\mathbb{1}_A \otimes (U - U'))(A \otimes \mathbb{1}_B)$ and hence $0 = (\mathbb{1}_A \otimes (U - U'))(X_A \otimes \mathbb{1}_{FB})(A \otimes \mathbb{1}_B)|\psi\rangle$. By linearity, this implies $U - U' = 0$.
- Step 12 computes the operator K in the GKLS-form according to the discussion above.
- Step 13 defines an operator K_{sc}, which according the statement of Theorem V.6 and also due to the discussion below Eq. (16) is of the form $K_A \otimes \mathbb{1}_B + \mathbb{1}_A \otimes iH_B$.
- Steps 14 and 15 extract K_A and H_B from K_{sc}. Note that such a decomposition is not unique, since for any $\lambda \in \mathbb{R}$, the transformation $K_A \to K_A + i\lambda\mathbb{1}_A$, $H_B \to H_B - \lambda\mathbb{1}_B$ leaves K_{sc} invariant. This transformation, however, allows us to choose H_B traceless. In that case, steps 14 and 15 determine K_A and H_B.

APPENDIX D: CONSTRUCTIVE APPROACH TO THEOREM V.18

In this appendix, we are going to describe in detail how one can computationally construct the operators A, U, B, H_A, and H_B in Theorem V.18 if the conditions of Lemma V.17 are met. We use the notation from Appendix C.

Given the operator $\hat{\mathfrak{L}} \in \mathcal{B}(\mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2})$ as in Lemma V.17, then we can compute the operators A, U, B, H_A, and H_B via the following eight steps:

1. Apply steps 1–18 in the protocol in Appendix C to $\hat{\mathfrak{L}}$. This yields $\mathcal{H}_E = \mathcal{H}_{A_1} \otimes \mathcal{H}_{B_1} \otimes \mathcal{H}_{A_2} \otimes \mathcal{H}_{B_2}$, $\tilde{A} \in \mathcal{B}(\mathcal{H}_{A_2}; \mathcal{H}_{A_0} \otimes \mathcal{H}_E)$, $\tilde{U} \in \mathcal{B}(\mathcal{H}_E \otimes \mathcal{H}_B; \mathcal{H}_B \otimes \mathcal{H}_E)$, $\tilde{K}_A \in \mathcal{B}(\mathcal{H}_A)$, and $\tilde{H}_B \in \mathcal{B}(\mathcal{H}_B)$.
2. Choose any unit vector $|\xi\rangle \in \mathcal{H}_E$.
3. Compute $\sigma = |\xi\rangle\langle\xi|$.
4. Compute $A = (\mathbb{1}_{A_{-1}} \otimes \mathbb{1}_E \otimes \langle\Omega_{A_0;A_3}|)(\mathbb{1}_{A_{-1}} \otimes \mathbb{F}_{A_0;E}\tilde{A} \otimes \mathbb{1}_{A_3})(|\Omega_{A_{-1};A_2}\rangle \otimes \mathbb{1}_{A_3} \otimes \langle\xi|)$.
5. Compute $B = \tilde{B}(\mathbb{1}_B \otimes \langle\xi|)$.
6. Compute $U = \tilde{U}\mathbb{F}_{B;E}$.
7. Set $H_B = \tilde{H}_B$.
8. Calculate $H_A = \frac{1}{2i}(\tilde{K}_A^T - \tilde{K}_A^{\dagger T})$, where the transposition is with respect to the $\{|a_i\rangle\}$ basis defined in Appendix C.

Let us comment on why the steps above yield the right result:

- Step 1 can be executed, since the assumptions of Lemma V.5 are the first two assumptions in Lemma V.17.
- Steps 2 and 3 define σ as in the Proof of Theorem V.18.
- Step 4 is a more explicit expression for $\tilde{A}^{T_A}\Xi_A^\dagger$ in the Proof of Theorem V.18.
- Steps 5, 6, and 7 are exactly the definitions of B, U, and $\mathcal{H}_B$, respectively, in the Proof of Theorem V.18.
- For step 8, we note that the condition $\mathrm{tr}_{A_1}[\hat{\mathfrak{L}}^A] = 0$ implies $L^A(\mathbb{1}) = 0$ so that we can follow the last few sentences in the Proof of Theorem V.18.

REFERENCES

[1] G. Chiribella, G. M. D'Ariano, and P. Perinotti, Europhys. Lett. **83**, 30004 (2008).
[2] D. Beckman, D. Gottesman, M. A. Nielsen, and J. Preskill, Phys. Rev. A **64**, 052309 (2001).
[3] G. Chiribella, G. M. D'Ariano, and P. Perinotti, Phys. Rev. Lett. **101**, 060401 (2008).
[4] G. Chiribella, G. M. D'Ariano, and P. Perinotti, Phys. Rev. A **80**, 022339 (2009).
[5] A. Bisio and P. Perinotti, Proc. R. Soc. London, Ser. A **475**, 20180706 (2019).
[6] O. Oreshkov, F. Costa, and Č. Brukner, Nat. Commun. **3**, 1092 (2012).
[7] G. Chiribella, G. M. D'Ariano, P. Perinotti, and B. Valiron, Phys. Rev. A **88**, 022318 (2013).
[8] E. Castro-Ruiz, F. Giacomini, and Č. Brukner, Phys. Rev. X **8**, 011047 (2018).
[9] J. H. Selby, A. B. Sainz, and P. Horodecki, "Revisiting dynamics of quantum causal structures—When can causal order evolve?," arXiv:2008.12757 [quant-ph] (2020).
[10] T. Eggeling, D. Schlingemann, and R. F. Werner, Europhys. Lett. **57**, 782 (2002).

[11] M. Piani, M. Horodecki, P. Horodecki, and R. Horodecki, Phys. Rev. A **74**, 012305 (2006).
[12] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Rev. ed. (Academic Press, San Diego, CA, 1980).
[13] S. Attal, Tensor products and partial traces, 2021.
[14] E. Davies, *Quantum Theory of Open Systems* (Academic Press, 1976).
[15] S. Attal, Quantum channels, 2021.
[16] M.-D. Choi, Linear Algebra Appl. **10**, 285 (1975).
[17] A. Jamiołkowski, Rep. Math. Phys. **3**, 275 (1972).
[18] K. J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Graduate Texts in Mathematics (Springer, New York, 2006).
[19] G. Lindblad, Commun. Math. Phys. **48**, 119 (1976).
[20] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan, J. Math. Phys. **17**, 821 (1976).
[21] M. M. Wolf, J. Eisert, T. S. Cubitt, and J. I. Cirac, Phys. Rev. Lett. **101**, 150402 (2008).
[22] D. Evans and J. Lewis, *Dilations of Irreversible Evolutions in Algebraic Quantum Theory*, Communications of the Dublin Institute for Advanced Studies, Series A Vol. 24 (Dublin Institute for Advanced Studies, 1977).
[23] A. Bátkai, M. Fijavž, and A. Rhandi, *Positive Operator Semigroups: From Finite to Infinite Dimensions*, Operator Theory: Advances and Applications (Springer International Publishing, 2017).
[24] T. M. Liggett, *Continuous Time Markov Processes: An Introduction* (American Mathematical Society, 2010), Vol. 113.
[25] G. Chiribella, A. Toigo, and V. Umanità, Open Syst. Inf. Dyn. **20**, 1350003 (2013).
[26] D. Kretschmann and R. F. Werner, Phys. Rev. A **72**, 062323 (2005).
[27] B. Collins, Int. Math. Res. Not. **2003**, 953.
[28] B. Collins and P. Śniady, Commun. Math. Phys. **264**, 773 (2006); arXiv:0402073 [math-ph].
[29] M. Fukuda, R. König, and I. Nechita, J. Phys. A: Math. Theor. **52**, 425303 (2019).
[30] Uniqueness of such a limit is clear. Existence follows by the Banach–Alaoglu theorem and an application of the uniform boundedness principle, which implies that the sequence $\Phi_{m,n}^{A}(X_A)$ is norm-bounded.
[31] M. M. Wolf, Quantum channels and operations: Guided tour, 2012.
[32] The partial trace $\mathrm{tr}_{\mathbb{C}}[\cdot]$ over the one-dimensional space $\mathbb{C}$ is just to ensure formal similarity with Lemma V.20.
[33] G. Gour and M. M. Wilde, Phys. Rev. Res. **3**, 023096 (2021).
[34] G. Gour, IEEE Trans. Inf. Theory **65**, 5880 (2019).
[35] P. Faist, M. Berta, and F. Brandão, Phys. Rev. Lett. **122**, 200601 (2019).
[36] S. Chen and E. Chitambar, Quantum **4**, 299 (2020).
[37] T. S. Cubitt, J. Eisert, and M. M. Wolf, Commun. Math. Phys. **310**, 383 (2012).
[38] T. S. Cubitt, J. Eisert, and M. M. Wolf, Phys. Rev. Lett. **108**, 120503 (2012).
[39] E. Onorati, T. Kohler, and T. Cubitt, "Fitting quantum noise models to tomography data," arXiv:2103.17243 [quant-ph] (2021).
[40] M. M. Wolf and J. I. Cirac, Commun. Math. Phys. **279**, 147 (2008).
[41] J. Bausch and T. Cubitt, Linear Algebra Appl. **504**, 64 (2016).
[42] M. C. Caro and B. R. Graswald, J. Math. Phys. **62**, 042203 (2021).
[43] B. Nachtergaele, R. Sims, and A. Young, J. Math. Phys. **60**, 061101 (2019).