Technische Universität München

TUM School of Computation, Information and Technology

Reduced basis methods for problems with moving features

Tobias Michael Blickhan

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Daniel Matthes

Prüfende der Dissertation:

1. Prof. Dr. Eric Sonnendrücker
2. Prof. Dr. Tomasz Tyranowski
3. Prof. Dr. Virginie Ehrlacher

Die Dissertation wurde am 13.11.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 15.05.2024 angenommen.

Summary

This work deals with reduced order methods for the solution of parametrized partial differential equations with moving features. Solutions of these class of equations are well known to be hard to approximate using linear reduced basis methods. We utilize results from optimal transportation theory to extend these methods. In particular, we present a novel registration method that is based on aligning solution features using an approximation of optimal transport maps.

Zusammenfassung

Diese Arbeit beschäftigt sich mit Reduzierten Methoden zur Lösung von parameterabhängigen partiellen Differentialgleichungen, welche von Advektionseffekten dominiert werden. Es ist bekannt, dass Lösungen dieser Gleichungen nur schwer mit linearen Reduzierten Basis Methoden dargestellt werden können. Wir nutzen Ergebnisse aus der Theorie des Optimalen Transportes, um diese Methoden zu erweitern.

List of contributed articles

Tobias Blickhan
A registration method for reduced basis problems using linear optimal transport
SIAM Journal on Scientific Computing, 46(5) (2024), pp. A3177-A3204
https://doi.org/10.1137/23M157071

Beatrice Battisti, Tobias Blickhan, Guillaume Enchery, Virginie Ehrlacher, Damiano Lombardi and Olga Mula
Wasserstein model reduction approach for parametrized flow problems in porous media
ESAIM: Proceedings and Surveys, 73 (2023), pp. 28-47
https://doi.org/10.1051/proc/202373028

Contents

Chapter 1

Introduction

Numerical simulations are an indispensable tool in order to understand the behavior of complex physical systems. They are ubiquitous in many more fields of applications of science and engineering, used to gain insights on questions from economics to biology, from the scale of subatomic particles to the construction of skyscrapers or even cosmological structures.

The algorithms and frameworks employed (finite element methods, Monte Carlo methods, and, more recently, neural networks, to name a few) have seen a massive development since the inception of the first computers to solve mathematical problems around the time of the second world war. The number of simulation software grows daily and, since the available computing power also increased, the problems that can be tackled have grown from Monte Carlo simulations of hundreds of neutrons to simulations with over 500 billion particles and 28 billion grid points [59].

Next to experimental observation and theoretical modelling, numerical simulation is one of the three pillars of scientific work and engineering design processes. Naturally, simulation software is developed to carry out a large number of simulations. This is precisely the advantage of simulation-based design compared to the construction of prototypes and experimental setups.

1.1 Parametrized and many-query problems

In typical cases, simulation runs are performed for a number of parameter configurations. Parameters can encode initial and boundary conditions, geometrical configurations, or physical and material properties. Consider the case where the simulation solves a parameter-dependent partial differential equation (PPDE):

Definition 1.1 (Parametrized partial differential equation problem)**.** *Given a domain $\Omega \subset \mathbb{R}^d$, parameter $\mu \in \mathcal{A} \subset \mathbb{R}^p$, function space V, and $\mathcal{L} : V \times \mathcal{A} \to \mathbb{R}$, find $u(\mu) \in V$ such that*

$$\mathcal{L}(\mu, u(\mu)) = 0. \tag{1.1}$$

The operator $\mathcal{L}$ includes the differential operators, boundary conditions, and forcing terms of the problem.

Example 1.1 (Optimization)**.** *Consider the task of finding a particular parameter value minimizing some functional $\min_{\mu \in \mathcal{A}} J(\mu, u(\mu))$ where $u(\mu)$ is defined by*

Equation (1.1)*. It is clear that any form of iterative method to minimize* $J(\mu, u(\mu))$ *requires solving Equation* (1.1) *numerous times for different* μ*.*

Example 1.2 (Inverse problems)**.** *Consider the case where* $u^{\text{obs}} := u(\mu^{\text{obs}})$ *is known (for example through measurements) and we want to find the corresponding* μ^{obs}*. After adequate regularization (the data is typically noisy and we have no knowledge about the continuity of* $u(\mu) \mapsto \mu$*), this can be written as a PDE-constrained optimization problem.*

Practical considerations may require these problems to be solved very quickly, possibly in real time, and with limited computational resources. Consider, for example, that $J(u(\mu))$ is a security-relevant diagnostic of a power plant and μ are parameters of operation. These cases fall into what is commonly referred to as *many-query* context.

This setting adds additional challenges to the solution method for PPDE problem (1.1). While a classical high-fidelity numerical method such as a highly resolved finite element method is adequate to solve Equation (1.1) once, placing the same method inside an optimization loop will mean that it gets called hundreds of times. This drastically increases the computational time and energy consumption. In the case of embedded systems or real-time applications, running the high fidelity model once can already be infeasible. Furthermore, storing the results of numerous simulations in high resolution can lead to extreme memory requirements.

1.2 Reduced order modelling

The goal of reduced order models (ROMs) is to build a computationally cheap, yet sufficiently accurate approximation of the map $\mu \mapsto J(\mu, u(\mu))$ in the case of Example 1.1, or $u^{\text{obs}} \mapsto \mu^{\text{obs}}$ in the case of Example 1.2, or $\mu \mapsto u(\mu)$ in general. One might wonder if such an approximation even exists, given that the solution of Equation (1.1) typically requires the use of sophisticated numerical methods that approximate $u(\mu)$ with a large number N of discrete degrees of freedom. Even when J and $\mathcal{L}(\cdot; \mu)$ are linear (in u or even in both arguments), this map is in general non-linear.

At the same time, the dimension p of $\mathcal{A} \ni \mu = (\mu_1, \ldots, \mu_p)$ is often moderate, and the map $\mu \mapsto u(\mu)$ is regular. For example, in the case of elliptic PDEs (uniformly with respect to μ), the latter can be rigorously proved as we will discuss in Chapter 2.

General and tailored approximation spaces

General numerical approaches to solve PDEs do so by approximating elements of function spaces V using elements of large, finite dimensional vector spaces. In the case of an elliptic problem, V is usually given by the space H^1, comprised of square integrable functions with square integrable (weak) derivative.

In order to achieve a satisfactory approximation to this space, high resolution and a large number N of degrees of freedom is necessary. Ultimately, commonly used numerical methods, when applied to Equation (1.1), lead to a large system of equations that stems from the discretization of Equation (1.1) and requires the inversion of large (albeit sparse) $N \times N$ matrices.

Remark 1.1. *In many application cases, the solution $u(\mu)$ is only needed as an intermediate step to evaluate $J(\mu, u(\mu))$. As such, both input μ and output $J(\mu, u(\mu))$ can be low-dimensional even if N is very large.*

When solving several instances of a PPDE as in Example 1.1, one does not explore the entire space V but only a small (in a sense to be made more precise in Definition 2.1) fraction of it, namely those elements of V that are solutions to Equation (1.1) for some $\mu \in \mathcal{A}$. This set we refer to as the *solution manifold* of the PPDE problem:

Definition 1.2 (Solution manifold)**.** *The solution manifold of the PPDE problem defined by Equation* (1.1) *is given by*

$$\mathcal{M} := \{u(\mu) \in V : \mathcal{L}(u(\mu); \mu) = 0 \textit{ where } \mu \in \mathcal{A}\}. \tag{1.2}$$

Remark 1.2. *The term* manifold *is used in the literature interchangeably with* solution set*. In particular, it is used without verifying that $\mathcal{M}$ is in fact locally homeomorphic to Euclidean space.*

For the sake of multi-query problems, the numerical method employed only has to be able to approximate elements of $\mathcal{M}$.

Remark 1.3. *At first glance, it seems that one can use the parameters $\mu_1, \ldots, \mu_p$ as coordinates on $\mathcal{M}$ through the functions $\mu \to u(\mu)$ and that p is an indicator for how hard $\mathcal{M}$ is to approximate.*

This can be misleading in two ways. The map $\mu \to u(\mu)$ can be discontinuous or very irregular. As pointed out in [18], it can have properties similar to those of a space-filling curve and require a large number of (more regular) functions to approximate.

On the other hand, the parametrization of the problem can be redundant and the dimension of $\mathcal{M}$ in fact much smaller than p. In practice, it is often the case that even if a physical system is described by hundreds of parameters, an experienced domain scientist can make reliable heuristic predictions based on the knowledge of only a few variables. We take this as an indication that the intrinsic dimension of the system is in fact much smaller.

Offline-online splitting

The field of reduced complexity modelling is of practical interest for numerous multi-query and real-time applications. In this work, we investigate methods with an *offline-online-splitting.*

First, the reduced model is constructed in the *offline or training phase*, leveraging the strengths of high-performance computing infrastructure. After construction, reduced models are evaluated at low computational cost in the *online phase*, reducing the time, money, and energy spent on optimization loops, inverse problems, or routine calculations.

Classical methods such as the reduced basis approach [108, 101] provide ways to reduce computational cost by orders of magnitude while at the same time ensuring rigorous error bounds for the reduced simulation of elliptic and parabolic equations. However, they are notoriously ill-suited when working with hyperbolic systems or solutions with moving features and sharp discontinuities as will be discussed in Section 2.6.

1.3 Scope of this thesis

This thesis covers reduced order modelling, a topic on the intersection of data science and scientific computing. We draw heavily on concepts from optimal transport theory, which establishes connections to convex analysis.

The presentation is targeted at an audience with a background in scientific computing. An effort has been made to keep the presentation of most concepts from optimal transport theory and reduced complexity modelling self-contained, therefore the content of important results, algorithms, and theorems is included. Whenever possible, these results are accompanied by proofs, possibly in a less general, simplified setting, or at least formal considerations that make them plausible. When it seems unfeasible to do so, and for an exhaustive treatment of the theory in general, we refer to the provided references.

The thesis is structured as follows: Chapter 2 introduces the reduced basis method as well as registration methods in particular. Results from optimal transport theory are covered in Chapter 3, while Chapter 4 introduces results and algorithms used to compute it. Chapter 5 discusses existing and new concepts of how to leverage optimal transport theory for reduced complexity modelling of partial differential equations that are characterized by moving features or can be formulated naturally on the space of probability densities. In Chapter 6, we introduce a novel method to build a registration-based reduced basis method using the ideas presented so far. Connections between optimal transport theory and variational models in fluid mechanics that are not directly related to the reduced order modelling, yet interesting especially for plasma physics applications, are discussed in Appendix B.

Our main contributions to the field are presented in Section 5.4 and Chapter 6. The results of Section 5.4 were largely produced during a stay of the author at the CEMRACS event of 2021 in collaboration with Beatrice Battisti, Guillaume Enchéry, Virginie Ehrlacher, Damiano Lombardi, and Olga Mula.

The numerical results presented throughout this work have been obtained using codes developed by the author which are available at `https://github.com/ToBlick`.

1.4 Notation

We briefly go over our notational choices. A table of commonly used symbols, operators, and abbreviations is also given in Appendix A.

Functions, spaces, function spaces

Unless explicitly stated otherwise, Ω will denote a subset of $\mathbb{R}^d$, $d \in \mathbb{N} = \{1, 2, \dots\}$. $\mathbb{R}_{>0}$ denotes strictly positive elements of $\mathbb{R}$.

The closure of Ω, denoted $\overline{\Omega}$, is the smallest closed set containing Ω. The boundary of Ω is denoted by $\partial\Omega$ and is the difference between its closure and interior, the latter is the smallest open set that contains Ω.

The space of $\mathbb{R}^d$-valued continuous functions on Ω will be denoted by $\mathcal{C}(\Omega, \mathbb{R}^d)$ and equipped with the supremum norm. The space of continuous functions valued in $\mathbb{R}$ is denoted $\mathcal{C}(\Omega)$. The space of continuous functions on Ω that vanish on $\partial\Omega$ is denoted $\mathcal{C}_0(\Omega)$ and that of bounded continuous functions is denoted $\mathcal{C}_b(\Omega)$. The space $C^k(\Omega)$ for $k \in \mathbb{N} \cup \{0, +\infty\}$ consists of continuous functions on Ω with k continuous derivatives and $C^{0,\alpha} : \alpha \in (0, 1)$ denotes Hölder continuous functions, i.e. those f that satisfy

$$|f(x) - f(y)| \le C|x - y|^\alpha \tag{1.3}$$

for all $x, y \in \Omega$ and some $C > 0$. The space $C^{k,\alpha}$ is the subset of C^k with ($\alpha-$)Hölder continuous kth partial derivatives. If $f \in C^{0,1}$, it is Lipschitz continuous. The special case when $C = 1$ in Equation (1.3) we call 1-Lipschitz.

We denote by $L^p(\Omega, \rho, \mathbb{R}^d)$ the space of functions $u : \Omega \to \mathbb{R}^d$ such that

$$\|u\|^p_{L^2(\rho)} := \int |u|^p \,\mathrm{d}\rho < +\infty. \tag{1.4}$$

For the sake of brevity, we will also use the shorthand notation $L^p(\rho)$, as in the definition of the norm above.

V will denote a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$. Examples are $L^2(\Omega)$, the space of square integrable functions on Ω with square integrable weak first derivative, denoted $H^1(\Omega)$, and the closure of $C_0^\infty(\Omega)$ in $H^1(\Omega)$, denoted $H^1_0(\Omega)$. Again for the sake of brevity, we write $\|\cdot\|_{L^2}, \|\cdot\|_{H^1}, \dots$ for the corresponding norms. The notation $\|u\|^2_{\dot{H}^1} := \int \nabla u(x) \cdot \nabla u(x) \,\mathrm{d}x$ denotes the squared H^1 semi-norm.

We reserve this notation for the inner product and norms to function spaces. The inner product in $\mathbb{R}^d$ will be denoted $x \cdot y$ and the Euclidean norm $|x|$.

When considering functions with parameter dependence $u : \mathcal{A} \times \Omega \to \mathbb{R}$ valued $u(\mu, x)$, we will often write u_μ to refer to $u(\mu, \cdot) : \Omega \to \mathbb{R}$.

Measures

A probability measure $\rho \in \mathcal{P}(\Omega)$ assigns a positive number $\rho[\Omega']$ to any measurable subset of Ω where $\rho[\Omega] = 1$. As already stated, we denote the integration of a measurable function f by $\rho(f) = \int f \,\mathrm{d}\rho$ or $\int_\Omega f(x) \,\mathrm{d}\rho(x)$ for additional clarity. We mention but do not use the equivalent notations $\langle f, \rho \rangle$ and $\mathbb{E}_{X \sim \mu}[f(X)]$. The support $\operatorname{supp}(\rho)$ of a measure ρ is defined as the smallest closed set such that $\rho[\Omega \setminus \operatorname{supp}(\rho)] = 0$.

The (d-dimensional) Lebesgue measure of a set Ω will be denoted $|\Omega|$.

When we say that a measure is absolutely continuous, we always mean with respect to the Lebesgue measure unless explicitly stated otherwise. If ρ is absolutely continuous, we write $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and we will call it a density. In order to avoid introducing additional notation and by abuse of notation, we will call the density of ρ with respect to the Lebesgue measure ρ as well.

The Dirac measure at x will be denoted δ_x and is defined by $\int \varphi \delta_x = \varphi(x) \; \forall \varphi \in \mathcal{C}(\Omega)$.

The *narrow* convergence of measures $\in \mathcal{P}(\Omega)$ in duality with $\mathcal{C}_b(\Omega)$ is denoted by $\rho_n \rightharpoonup \rho \Leftrightarrow \int \varphi \mathrm{d}\rho_n \to \int \varphi \mathrm{d}\rho \; \forall \varphi \in \mathcal{C}_b(\Omega)$. Note that narrow convergence coincides with weak convergence in duality with $\mathcal{C}(\Omega)$ in the case where Ω is compact.

Chapter 2

The reduced basis method

In this work, we will employ the reduced basis (RB) approach for parametrized partial differential equations. Standard references on this topic that we will use and refer to are [108, 18]. Recall the PPDE problem from Definition 1.1: Given parameters $\mu \in \mathcal{A} \subset \mathbb{R}^p$, we seek to solve many iterations of the following problem: Find a function $u(\mu) \in V$, a Hilbert space, such that

$$\mathcal{L}(\mu, u(\mu)) = 0 \quad \text{in } \Omega.$$

To solve this problem numerically, classical numerical methods such as the finite difference, finite element, spectral, or finite volume methods approximate the space V using a high-dimensional vector space V_h ($\dim V_h = N$). Elements of V_h, denoted by u_h, are determined by a degree of freedom vector $\mathrm{u} \in \mathbb{R}^N$. Using a suited discretization of $\mathcal{L}$ denoted $\mathcal{L}_h$ (in the conforming case $V_h \subset V$, one can use the restriction of $\mathcal{L}$ to V_h), one obtains the discretized system $\mathcal{L}_h(\mu, u_h(\mu)) = 0$. In general, this corresponds to a very large non-linear system of equations for $\mathtt{u}$.

2.1 n-width

Recall the definition of the solution manifold

$$\mathcal{M} := \{u(\mu) \in V : \mathcal{L}(u(\mu); \mu) = 0 \text{ where } \mu \in \mathcal{A}\}.$$

As mentioned in Chapter 1, one approach in reduced order modelling is to build a tailored, low-dimensional approximation space for elements of $\mathcal{M}$ rather than general elements of V. In the case of the reduced basis method, this is a linear subspace that we will denote V_n ($\dim V_n = n$). How the approximation error decays as the dimension grows is a property of $\mathcal{M}$ that we call *linear compressibility*[1]. This is formalized with the concept of *n-width* ([108], Section 5.4):

Definition 2.1 (n-width)**.** *The* (Kolmogorov) n-width *of $\mathcal{M}$ from Definition 1.2 is given by*

$$d_n(\mathcal{M}, V_h) := \inf_{\substack{V_n \subset V_h \\ \dim V_n = n}} \sup_{\mu \in \mathcal{A}} \inf_{u_{\mathrm{rb}} \in V_n} \|u_{\mathrm{rb}}(\mu) - u(\mu)\|_V. \tag{2.1}$$

[1] Compressibility is used here to refer to the succinct description of data and has little to do with divergence-free vector fields in fluid dynamics.

The n-width is a worst-case error indicator: the distance between V and V_n is measured by the worst case (with respect to μ) of the best approximation of $u(\mu)$ by $u_{\mathrm{rb}}(\mu)$. Since V is a Hilbert space, there exists an orthogonal projection $\mathrm{Proj}_n : V \to V_n$ onto the reduced space.[2] We can therefore also write

$$d_n(\mathcal{M}, V_h) = \inf_{\substack{V_n \subset V_h \\ \dim V_n = n}} \sup_{\mu \in \mathcal{A}} \|\mathrm{Proj}_n u(\mu) - u(\mu)\|_V. \tag{2.2}$$

Instead of a worst-case error, we can also consider an average:

$$\delta_n^2(\mathcal{M}, V_h) := \inf_{\substack{V_n \subset V_h \\ \dim V_n = n}} \int_{\mathcal{A}} \|\mathrm{Proj}_n u(\mu) - u(\mu)\|_V^2 \mathrm{d}\mu. \tag{2.3}$$

Note that $\delta_n(\mathcal{M}, V_h) \leq \left(\int_{\mathcal{A}} \mathrm{d}\mu\right)^{1/2} d_n(\mathcal{M}, V_h)$. One can extend this definition by introducing a distribution on $\mathcal{A}$ that weighs the error according to the spread of parameter values for the problem at hand.

Example 2.1. *The n-width of the solution manifold consisting of $u \in H_0^1(\Omega)$ solving*

$$-\nabla \cdot (a(x;\mu)\nabla u(x;\mu)) = f(x;\mu) \quad \textit{in } \Omega \tag{2.4}$$

where $a(x;\mu)$ is bounded from below by a positive function, uniformly in μ, decays exponentially ([108], Section 5.5).

A very good reference for the notions of reducibility of solution sets as well as non-linear extensions of the n-width is Chapter 3 in [18].

2.2 Proper orthogonal decomposition

So far, our presentation concerned the continuous picture, where u are elements of a Hilbert space and a continuous distribution of parameters is available. In fact, many of the following considerations can be done while staying in this framework.

In practice, however, we are already starting from a finite-dimensional approximation of $u(\mu)$ given by $u_h(\mu)$, which is, for example, an element of a finite element space. We refer to $u_h(\mu)$ as the *high fidelity solution.* When discussing the error of a reduced method, we usually mean the deviation from this high fidelity solution, since we do not have access to $u(\mu)$. In the interest of staying close to the application and the numerical examples, we will from now on stay in this finite-dimensional setting.

Compression of degree of freedom data

Furthermore, assume the parameter space is sampled at n_s points $\mu_i, \ldots, \mu_{n_s}$, giving us a set of *snapshots* $u_h(\mu_i) \in \mathcal{M}_h$, where $\mathcal{M}_h$ is the solution manifold of the discrete full-order problem. Note that the sampling need not be uniform, this implies the introduction of a weighting of the integral over $\mathcal{A}$. We do not consider this case here,

[2] If $u_{\mathrm{rb}}^* = \arg\min_{u_{\mathrm{rb}} \in V_n} \|u_{\mathrm{rb}} - u\|_V$, then $\|u_{\mathrm{rb}} - u\|_V^2 \leq \|u_{\mathrm{rb}} + \epsilon v - u\|_V^2 \Leftrightarrow 2\epsilon\langle u_{\mathrm{rb}} - u, v\rangle \leq \epsilon^2 \|v\|_V^2$ for any $v \in V_n, \epsilon \in \mathbb{R}$ which implies the orthogonality $u_{\mathrm{rb}} - u \perp V_n$.

but introducing it in the following expressions is straightforward. In this setting, Equation (2.3) reads

$$\min_{\substack{V_n \subset V_h \\ \dim V_n = n}} \frac{1}{n_s} \sum_{i=1}^{n_s} \| \mathrm{Proj}_n u_h(\mu_i) - u_h(\mu_i) \|_{V_h}^2 . \tag{2.5}$$

On the level of degrees of freedom, $\langle u_h, v_h \rangle_{V_h} = \mathtt{u}^T \mathbb{M} \mathtt{v}$ with mass matrix $\mathbb{M}$ (symmetric, positive definite). Furthermore, an orthogonal (with respect to $\langle \cdot, \cdot \rangle_{V_h}$) projection Proj_n has a matrix representation as $\mathbb{W}\mathbb{W}^T\mathbb{M}$ where the columns of $\mathbb{W} \in \mathbb{R}^{N \times n}$ are $\mathbb{M}$-orthonormal, i.e. $\mathbb{W}^T\mathbb{M}\mathbb{W} = \mathrm{Id}_n$.

The solution to this problem is given by the Schmidt-Eckart-Young (also known as Eckart-Young-Mirsky) Theorem ([108], Theorem 6.1): The optimal matrix is obtained from a singular value decomposition of the snapshot matrix with elements $\sum_{k=1}^{N} \mathbb{M}_{ik}^{1/2}{}_k(\mu_j)$, where $1 \leq i \leq N$ and $1 \leq j \leq n_s$. In particular, a basis of the optimal subspace is given by the left eigenvectors of the snapshot matrix, scaled with $\mathbb{M}^{-1/2}$.

In our application cases, $N \gg n_s$, and it is convenient to obtain the basis from the *correlation matrix* $\mathbb{C}^u$:

Definition 2.2 (Snapshot correlation matrix)**.** *Given snapshots* $\{u_h(\mu_j)\}_{j=1}^{n_s} \subset \mathcal{M}_h$*, the elements of the snapshot correlation matrix are given by*

$$\mathbb{C}_{ij}^u := \langle u_h(\mu_i), u_h(\mu_j) \rangle_{V_h} \quad 1 \leq i, j \leq n_s. \tag{2.6}$$

Obtaining the POD basis is described in Algorithm 1.

Algorithm 1 POD algorithm

```
function PODBASIS($\{u_h(\mu_1), \dots, u_h(\mu_{n_s})\}, \tau$)
    for $i = 1, \dots, n_s$ and $j = 1, \dots, n_s$ do
        $\mathbb{C}_{ij}^u \leftarrow \langle u_h(\mu_i), u_h(\mu_j) \rangle_{V_h}$
    end for
    for $i = 1, \dots, n_s$ do
        $\lambda_i^u, \mathtt{f}_i^u \leftarrow \mathrm{EVD}(\mathbb{C}^u)$        ▷ $\mathbb{C}^u \mathtt{f}_i^u = \lambda_i^u \mathtt{f}_i^u$ such that $\lambda_1^u \geq, \lambda_2^u \geq \dots$.
        $\zeta_i \leftarrow (\lambda_i^u)^{-1/2} \sum_{j=1}^{n_s} u_h(\mu_j)(\mathtt{f}_i^u)_j$        ▷ $(\mathtt{f}_i^u)_j$ is $(\mathtt{f}_i^u)$s $j$th component.
    end for
    $n \leftarrow \min_{n'} : \sum_{i=1}^{n'} \lambda_i^u > (1 - \tau) \sum_{j=1}^{n_s} \lambda_j^u$
    return $\{\zeta_1, \dots \zeta_n\}$        ▷ The reduced basis
end function
```

The resulting basis enjoys the following optimality property, a direct consequence of the Schmidt-Eckart-Young Theorem.

Proposition 2.1 ([108], Proposition 6.2)**.** *The POD basis constructed by Algorithm 1 spans the optimal* n*-dimensional subspace in the sense of Equation* (2.5) *and the approximation error is given by*

$$\sum_{i=1}^{n_s} \left\| \sum_{j=1}^{n} \langle \zeta_j, u_h(\mu_i) \rangle_{V_h} \zeta_j - u_h(\mu_i) \right\|_{V_h}^2 = \sum_{i=n+1}^{n_s} \lambda_i^u . \tag{2.7}$$

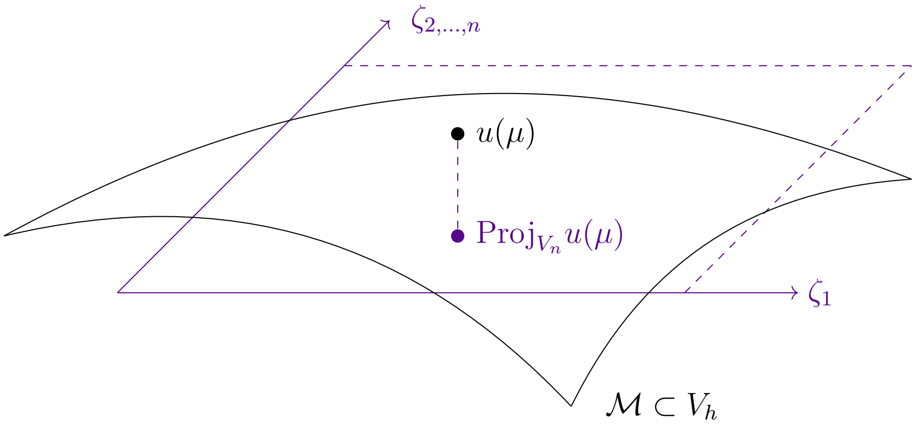

Figure 2.1: Illustration of the reduced basis approach.

When we consider a reduced basis approximation $\in \mathrm{span}\{\zeta_i\}_{i=1}^n$, we denote its coefficients by $\tilde{u}_i : i = 1, \dots, n$, i.e.

$$u_{\mathrm{rb}}(\mu, x) := \sum_{i=1}^{n} \tilde{u}_i(\mu)\zeta_i(x). \tag{2.8}$$

Generalizations

As mentioned, we usually rely on an eigenvalue decomposition of the $n_s \times n_s$ correlation matrix $\mathbb{C}^u$ rather than the $N \times n_s$ snapshot matrix. The notation we employed suggests that as long as we stay in a separable Hilbert space setting where we can work with orthogonality and bases, analogous considerations hold, and this is indeed the case.

Remark 2.1. *In a semi-discrete setting, we consider $u_h(\mu)$ an element of $L^2(\mathcal{A}, V_h)$. In this case, the snapshot correlation matrix $\mathbb{C}^u$ is replaced by the operator*

$$\mathbb{C}^u_{L^2(\mathcal{A},V_h)} : L^2(\mathcal{A}) \to L^2(\mathcal{A}) : g \mapsto \int_{\mathcal{A}} \langle u_h(\mu), u_h(\mu') \rangle_{V_h} g(\mu') \mathrm{d}\mu'. \tag{2.9}$$

$\mathbb{C}^u_{L^2(\mathcal{A},V_h)}$ shares eigenvalues with the operator

$$V_h \ni v \mapsto \int_{\mathcal{A}} \langle u_h(\mu), v \rangle_{V_h} u_h(\mu) \mathrm{d}\mu \in V_h, \tag{2.10}$$

which is of finite rank $\leq N$ and bounded, as $u_h(\mu) \in L^2(\mathcal{A}, V_h)$. Hence,

$$\int_{\mathcal{A}} \langle u_h(\mu), v \rangle_{V_h} u_h(\mu) \mathrm{d}\mu \leq \|u_h(\mu)\|^2_{L^2(\mathcal{A},V_h)} \|v\|_{V_h}. \tag{2.11}$$

Any bounded finite-rank operator between Hilbert spaces is compact[3] *and allows an eigendecomposition of the form*

$$\mathbb{C}^u_{L^2(\mathcal{A},V_h)}(g) = \sum_i \lambda^u_i \int_{\mathcal{A}} f^u_i(\mu') g(\mu') \mathrm{d}\mu' \, f^u_i(\mu), \tag{2.12}$$

[3] We call an operator between Hilbert spaces U and V is compact if its image of any bounded sequence in U has a convergent subsequence in V.

where $\{f_i^u\}_i \subset L^2(\mathcal{A})$ are orthonormal. This setting is discussed in [108], Section 6.4.

Remark 2.2. *When V_h is replaced by a general Hilbert space V, Equation* (2.12) *is still valid, as $\mathbb{C}^u_{L^2(\mathcal{A},V)}$ is a Hilbert-Schmidt integral operator. In the present case where $L^2(\mathcal{A})$ is separable, $u(\mu) \in L^2(\mathcal{A}, V)$, and $\mu \mapsto \langle u(\mu'), u(\mu)\rangle_V$ is in $L^2(\mathcal{A})$ for all $\mu' \in \mathcal{A}$, it is in fact* trace-class. *Its eigenvalues are at most countably infinite, real, and converge to zero. The sequence $\{\lambda_i^u\}_i$ is summable and*

$$\|\mathbb{C}^u_{L^2(\mathcal{A},V)}\|^2_{HS} = \sum_i \lambda_i^u = \int_{\mathcal{A}} \langle u(\mu), u(\mu)\rangle_V \,\mathrm{d}\mu. \tag{2.13}$$

For definitions and proofs we refer to [43] and [127].

Eigenvalue criteria

The eigenvalue decay of $\mathbb{C}^u$ is a good indicator of the best average approximation $\delta_n(\mathcal{M}, V_h)$ as long as the error introduced by approximating the integral over $\mathcal{A}$ is small compared to the approximation error itself. A common strategy to pick the right size n of a reduced basis is based on a user-defined tolerance τ and an energy criterion, as already shown in Algorithm 1.

Definition 2.3 (Eigenvalue energy criterion)**.** *Given the snapshot correlation matrix $\mathbb{C}^u \in \mathbb{R}^{n_s \times n_s}$ with eigenvalues $\lambda_1^u \geq \lambda_2^u \geq \dots$ and a tolerance τ, the size of the POD basis $n = n(\tau)$ is given by*

$$n := \arg\min \left(n' \in \mathbb{N} : \mathcal{E}(n', \{\lambda_i^u\}_i) := \frac{\sum_{i=1}^{n'} \lambda_i}{\sum_{j=1}^{n_s} \lambda_s} > 1 - \tau \right). \tag{2.14}$$

We call $\mathcal{E}(n, \{\lambda_i^u\}_i)$ the (n-dependent) *eigenvalue energy* of $\mathbb{C}^u$.

2.3 Greedy algorithms

Another strategy to determine a set of reduced basis elements $\zeta_1, \dots, \zeta_n$ relies on a greedy procedure: In an iterative process, elements are added to the basis that at this moment offer the biggest increase in approximation accuracy, measured in the L^∞ norm. A straightforward version of this is given in Algorithm 2.

In practice, the optimality criterion in Line 8 is replaced with an a posteriori error estimate that can be computed at low cost. If this is done, the greedy algorithm requires solving the full-order PDE problem only n times, while the error estimate is evaluated $\approx n \times n_s$ times. This allows one to choose much larger values of n_s compared to the POD case. This procedure is sometimes called a *weak greedy* approach.

Error estimates are an important feature of reduced basis methods. They are similar in nature to those employed in high fidelity methods. The residual of the approximate solution is used to infer information about the error compared to the true solution. RB methods come with reliable error estimates are often referred to as *certified.*

Algorithm 2 POD algorithm

```
function GREEDYBASIS($\{\mu_1, \ldots, \mu_{n_s}\}, \tau$)
    $\mu^{(1)} \leftarrow$ INITIAL($\mu_1, \ldots, \mu_{n_s}$)
    $n \leftarrow 1$
    $\Delta \leftarrow 2\tau$
    while $\Delta > \tau$ do
        $u_h(\mu^{(n)}) \leftarrow$ PDESOLVE($\mu^{(n)}$)
        $\zeta_n \leftarrow$ ORTHONORMALIZE($u_h(\mu^{(n)}), \{\zeta_i\}_{i=1}^{n-1}$)
        $\Delta, \mu^{(n+1)} \leftarrow \max_{\mu \in \{\mu_1,\ldots,\mu_{n_s}\}} \|u_h(\mu) - \mathrm{Proj}_{\zeta_1,\ldots,\zeta_n} u_h(\mu)\|_{V_h}$
        $n \leftarrow n + 1$
    end while
    return $\{\zeta_1, \ldots \zeta_n\}$                    ▷ The reduced basis
end function
```

We will come back to the computation of error indicators after we introduced reduced basis solutions in the next section.

The greedy method is a feasible way to find a suited approximation space in the sense of the Kolmogorov n-width d_n, (c.f. Equation (2.1)), in an iterative manner. Note, however, that the employed optimality criterion is local, as is characteristic of greedy optimization. At iteration n, ζ_n is chosen with $\zeta_1, \ldots, \zeta_{n-1}$ fixed.

Remark 2.3. *Exponential/algebraic n-width decay translates to exponential/algebraic decay of the RB approximation error when using a greedy method to construct the basis [32, 23].*

Remark 2.4. *Just as for the proper orthogonal decomposition, it is crucial that the samples $\{\mu_1, \ldots, \mu_{n_s}\} \subset \mathcal{A}$ are chosen in a way that represents the entire parameter space.*

2.4 Reduced basis solutions

We will illustrate the solution of a reduced basis problem using the following example.

Example 2.2 (Linear uniformly coercive PPDE problem)**.** *Let V be a Hilbert space, $\mathcal{A} \subset \mathbb{R}^p$, and $V_h \subset V$ finite-dimensional. $a(\cdot, \cdot; \mu) : V \times V \to \mathbb{R}$ is a bilinear form, symmetric, uniformly (in μ) continuous and uniformly coercive. $f(\cdot; \mu) : V \to \mathbb{R}$ is a continuous linear form. Find $u(\mu) \in V_h \subset V$ such that*

$$a_\mu(u_h(\mu), v_h) = f_\mu(v_h) \quad \forall v_h \in V_h. \tag{2.15}$$

By the Lax-Milgram Theorem and Céa's Lemma, solutions to this problem satisfy

$$\|u_h(\mu) - u(\mu)\|_V \leq \frac{c}{\alpha} \inf_{v_h \in V_h} \|u(\mu) - v_h\|_V, \tag{2.16}$$

where c and α denote the μ-independent bounds on the constants of continuity $c(\mu)$ and coercivity $\alpha(\mu)$ of $a(\mu, \cdot, \cdot)$. More details are given, for example, in Section 2.4 of [108].

Galerkin reduced basis approach

In this work, we focus on Galerkin-type reduced basis solutions. That is, both trial and test space are replaced by $V_n := \operatorname{span}\{\zeta_i\}_{i=1}^n$. Since $V_n \subset V_h \subset V$ in this case, the coercivity is preserved when moving from V_h to V_n, just as it was in the case of the full-order discretization:

$$\alpha_n(\mu) = \inf_{v_n \in V_n} \frac{a_\mu(v_n, v_n)}{\|v_n\|_V^2} \geq \inf_{v_h \in V_h} \frac{a_\mu(v_h, v_h)}{\|v_h\|_V^2} = \alpha_h(\mu) \geq \alpha > 0 \quad \forall \mu \in \mathcal{A}. \tag{2.17}$$

We define the RB solution $u_{\mathrm{rb}}(\mu) = u_n(\mu) \in V_n$ by

$$a_\mu(u_n(\mu), v_n) = f_\mu(v_n) \quad \forall v_n \in V_n. \tag{2.18}$$

Note that v_n is also an admissible test function for the continuous problem, therefore we immediately find

$$a_\mu(u_n(\mu) - u(\mu), v_n) = 0 \quad \forall v_n \in V_n. \tag{2.19}$$

Proposition 2.2. *For the setting of Example 2.2 and for all $\mu \in \mathcal{A}$, the RB approximation is optimal in the sense that*

$$\|u_n(\mu) - u(\mu)\|_V \leq \frac{c(\mu)}{\alpha(\mu)} \inf_{v_n \in V_n} \|v_n - u(\mu)\|_V. \tag{2.20}$$

Proof. For any $v_n \in V_n$,

$$\begin{aligned} \alpha(\mu)\|u_n(\mu) - u(\mu)\|_V^2 &\leq a_\mu(u_n(\mu) - u(\mu), u_n(\mu) - u(\mu)) && (2.21) \\ &= \underbrace{a_\mu(u_n(\mu) - u(\mu), u_n(\mu) - v_n)}_{=0} + a_\mu(u_n(\mu) - u(\mu), v_n - u(\mu)) && (2.22) \\ &\leq c(\mu)\|u_n(\mu) - u(\mu)\|_V \|v_n - u(\mu)\|_V. && (2.23) \end{aligned}$$

□

As $u_n(x;\mu) = \sum_{i=1}^n \tilde{u}_i(\mu)\zeta_i(x)$ with coefficients $\tilde{u}_i(\mu) \in \mathbb{R}$, in the case of a linear equation such as Equation (2.15),

$$\begin{aligned} &a(u_n(\mu), v_n; \mu) = f(v_n; \mu) \; \forall v_n \in V_n \\ &\qquad \Leftrightarrow \sum_{i=1}^n \tilde{u}_i(\mu) a_\mu(\zeta_i, \zeta_j) = f(\zeta_j; \mu) \; \forall j = 1, \ldots, n. \end{aligned} \tag{2.24}$$

Parameter separability

An important special case is when the parameter dependence of a and f is *separable*:

Definition 2.4 (Parameter separable form)**.** *The bilinear form $a_\mu : V \times V \to \mathbb{R}$ is* parameter separable *if there exist functions $\theta_q^a : \mathcal{A} \to \mathbb{R}$ and parameter-independent bilinear forms $a_q : V \times V \to \mathbb{R}$, with $1 \leq q \leq Q_a$ and*

$$a(\mu, \cdot, \cdot) = \sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q \quad \forall \mu \in \mathcal{A}. \tag{2.25}$$

Separability of the linear form $f(\mu, \cdot) : V \to \mathbb{R}$ is defined analogously.

Remark 2.5. *Equation* (2.25) *is also known as* affine parameter dependence *in the literature. The functions $\mu \mapsto \theta_q^a(\mu)$ can still be non-linear.*

Remark 2.6. *If the a_q are coercive with constants α_q, then a sufficient criterion for uniform coercivity of a_μ is $\sum_{q=1}^{Q_a} \theta_q^a(\mu)\alpha_q \geq \alpha_0 > 0$.*

If Definition 2.4 applies to the problem at hand, then Equation (2.24) can be written as

$$\sum_{i=1}^{n} \tilde{u}_i(\mu) \sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q(\zeta_i, \zeta_j) = \sum_{q=1}^{Q_f} \theta_q^f(\mu) f_q(\zeta_j) \quad \forall 1 \leq j \leq n. \tag{2.26}$$

As discussed in the introduction, one of the paradigms of reduced basis methods is to split computations into an offline phase, which can involve costly operations that in particular depend on the size of the full-order model N and an online phase, where the PPDE is solved for a new parameter value $\mu \in \mathcal{A}$ in a quick manner - independent of N. In the case of a linear, parameter separable equation, this is realized: The offline phase consists of constructing the reduced basis $\{\zeta_i\}_{i=1}^n$ and precomputing $a_q(\zeta_i, \zeta_j)$ and $f_{q'}(\zeta_j)$ for all $1 \leq q \leq Q_a, 1 \leq q' \leq Q_f$, and $1 \leq i, j \leq n$. In the online phase, given μ, we evaluate all $\theta_q^a(\mu)$ and $\theta_{q'}^f(\mu)$, assemble the matrix $\{\sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q(\zeta_i, \zeta_j)\}_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$ and vector $\{\sum_{q=1}^{Q_f} \theta_q^f(\mu) f_q(\zeta_j)\}_{1\leq j\leq n} \in \mathbb{R}^n$, and solve the resulting linear $n \times n$ system.

Remark 2.7. *The matrix $\{a(\zeta_i, \zeta_j; \mu)\}_{1\leq i,j\leq n}$ arising in the RB method plays the role the stiffness matrix from finite element (and finite volume, discontinuous Galerkin) methods. In contrast to the latter, however, it is dense and small rather than large and sparse.*

Remark 2.8. *For quadratic problems involving trilinear forms such as $b_\mu(u, v, w) = \int \mu(u \cdot \nabla v)w$, it is possible to pre-compute a third order tensor of the form*

$$\{b(\zeta_i, \zeta_j, \zeta_k)\}_{1\leq i,j,k\leq n}. \tag{2.27}$$

In principle, polynomial nonlinearities of higher order can be treated in this way as well, however this is feasible only when n is sufficiently small.

Remark 2.9. *For problems that satisfy a (uniform) inf-sup condition, which in the discrete form for two Hilbert spaces $V_h \subset V, W_h \subset W$ reads*

$$\inf_{v\in V_h} \sup_{w\in W_h} \frac{a_\mu(v, w)}{\|v\|_V \|w\|_W} = \beta_h > 0 \quad \forall \mu \in \mathcal{A}, \tag{2.28}$$

stability of the full-order model is not enough to deduce stability of a reduced order model even when $V_n = \operatorname{span}\{\zeta_i\}_{i=1}^n \subset V_h$ and $W_k = \operatorname{span}\{\chi_i\}_{i=1}^k \subset W_h$.

This is analogous to the situation when discretizing an inf-sup stable continuous problem. In this case, V_h and W_h have to be chosen such that Equation (2.28) *holds. A classical example of this is the Stokes equation, we refer in particular to Chapter 5 in [26].*

The reason why inf-sup stability might get lost is evidently because $\sup_{w\in W_k} \leq \sup_{w\in W_h}$ *as $W_k \subset W_h$. In the reduced basis setting, there is a computationally* *feasible option to restore it: Enrich the space W_k with those elements $s_\mu(v)$ such that $\langle s_\mu(v), w\rangle_W = a(\mu, v, w)$ $\forall v \in V_n$. We denote this enriched space $W^{s_\mu}_{k+n}$. These additional elements $s_\mu(v)$ are called* supremizer modes *and guarantee the existence of an element in $W^{s_\mu}_{k+n}$ that is as good as any element in W_h to satisfy Equation* (2.28)*. Furthermore, s_μ is a linear operation, hence it can be applied to the basis of V_n to construct*

$$W^{s_\mu}_{k+n} = W_k \cup \operatorname{span}\{s_\mu(\zeta_i)\}_{i=1}^n. \tag{2.29}$$

Since the infimum over the smaller space V_n is strictly greater than that over V_h, the inf-sup stability holds for the reduced system with constant $\beta_n \geq \beta_h$.

However, note that the reduced basis space $W^{s_\mu}_{k+n}$ is now parameter-dependent. This space can be replaced by one that is spanned by all supremizer modes of a fixed training set $\{s_{\mu_i} \circ v(\mu_i)\}_{i=1}^{n_s}$. While this does not guarantee the inf-sup stability, it leads to a μ-independent basis and is often sufficient in practice, see [108], Section 9.3.

Error estimates

Lastly, let us return to the question of how to estimate $\|u_{\text{rb}}(\mu) - u_h(\mu)\|_{V_h}$ when we only have access to the former. It holds that

$$a_\mu(u_h(\mu) - u_{\text{rb}}(\mu), v_h) = f_\mu(v_h) - a_\mu(u_{\text{rb}}(\mu), v_h) \quad \forall v_h \in V_h, \tag{2.30}$$

since $u_h(\mu)$ solves the PPDE from Example 2.2. The term on the right-hand side is the residual of the RB approximation evaluated on the high fidelity space:

$$V_h' \ni r_\mu : V_h \to \mathbb{R} : v_h \mapsto f_\mu(v_h) - a_\mu(u_{\text{rb}}(\mu), v_h), \tag{2.31}$$

where V_h' denotes the dual space of V_h. The residual only contains parameter-dependent forms and the reduced basis solution u_{rb}. Yet, since we can bound $a_\mu(u_h(\mu) - u_{\text{rb}}(\mu), v_h)$ from above by continuity and from below by coercivity, $\|r_\mu\|_{V_h'}$ allows us to control $\|u_{\text{rb}}(\mu) - u_h(\mu)\|_{V_h}$. If a_μ and f_μ are parameter-separable, evaluation of $\|r_\mu\|_{V_h'}$ can be a computationally cheap way to obtain the upper bound $\|u_{\text{rb}}(\mu) - u_h(\mu)\|_{V_h} \leq \|r_\mu\|_{V_h'}/\alpha(\mu)$.

2.5 Hyper-reduction

A natural question is what one does in the case where the problem at hand is not parameter separable. A priori, for a general function $a_h : V_h \times V_h \times \mathcal{A} \to \mathbb{R} : (u_h, v_h, \mu) \mapsto a_h(u_h, v_h; \mu)$, it is not possible to evaluate the quantity

$$a_h\left(\mu, \sum_{i=1} \tilde{u}_i(\mu)\zeta_i, \zeta_j\right), \tag{2.32}$$

which is needed to assemble the reduced system, independent of N. While the use of a reduced basis method in this case might still offer moderate computational cost reduction (for example, the assembly of the linear system might be N-dependent, but solving it is not), this is in most cases not satisfactory.

The mapping $\mu \mapsto a_h\left(\sum_{i=1} \tilde{u}_i(\mu)\zeta_i, \zeta_j; \mu\right)$ itself has a low-dimensional input $\mu \in \mathcal{A} \subset \mathbb{R}^p$ and output $\in \mathbb{R}$. *Hyper-reduction methods* are designed to provide an approximation of this mapping that does no longer depend on N. A number of hyper-reduction methods have been developed and are widely applied, among them *Gappy POD* [58, 37], *empirical quadrature* [145], and *empirical interpolation* [13, 39, 53] methods. We will describe the latter in detail as we will use it later.

Empirical interpolation

The empirical interpolation method (EIM) seeks to approximate the evaluation of a general parameter-dependent function $g : (x, \mu) \to g(x; \mu)$, where $g(\mu) := g(\cdot; \mu) \in L^\infty(\Omega)$ for all μ, using a separable form $g(x; \mu) \approx \sum_{q=1}^{Q_g} \theta_q^g(\mu) X_q^g(x)$. The coefficients $\{\theta_q^g(\mu)\}_q$ are calculated online by solving an interpolation problem:

$$g(x_q^{\text{eim}}; \mu) = \sum_{q'=1}^{Q} \theta_{q'}(\mu) X_{q'}(x_q^{\text{eim}}) \quad \forall 1 \le q \le Q. \tag{2.33}$$

The functions $\{X_q^g\}_q$ and points $\{x_q^{\text{eim}}\}_q$ are constructed in the offline phase based on instances of g from the training set. There exist two approaches, just as in the construction of the reduced basis $\{\zeta_i\}_i$: a greedy algorithm and one based on POD. Again, we will focus on the latter as it is used in the examples of Section 6.8 and refer to the references for the former.

The empirical interpolation method based on POD selects the interpolation functions $\{X_q^g\}_q$ starting from a POD basis obtained from the snapshot correlation matrix of g with elements

$$\mathbb{C}_{ij}^g = \langle g(\mu_i), g(\mu_j)\rangle \quad 1 \le i, j \le n_s. \tag{2.34}$$

The inner product can be chosen depending on the problem at hand, the L^2 inner product is a sensible choice to fall back on. When g is part of a PPDE problem that we seek to solve using a POD-RB approach, the functions $\{g(\mu_i)\}_{i=1}^{n_s}$ will be available from solving the full-order model to calculate the snapshots $\{u(\mu_i)\}_{i=1}^{n_s}$. In general, however, the samples of $\mathcal{A}$ used to construct the reduced basis and the EI functions need not agree. Proceeding as for the construction of the reduced basis, we obtain eigenfunctions $\{\Xi_q\}_{q=1}^{Q_g} \subset L^\infty$, where Q_g is determined by an energy criterion τ_{eim}. These eigenfunctions serve as the starting point to build the interpolation functions $\{X_q\}_q$, as described in Algorithm 3.

By construction, the resulting interpolation problem is solvable in $\mathcal{O}(Q^2)$ operations. The resulting approximation is always of parameter separable form and allows an evaluation in the online phase independent of N.

Remark 2.10. *Just as in the case of the POD construction of a reduced basis for solutions of a PPDE, the decay of eigenvalues of $\mathbb{C}^g$ gives a good indication of how well g can be approximated by empirical interpolation.*

Remark 2.11. *Oversampling techniques have been shown to improve the stability of the empirical interpolation method in the presence of noisy data [105].*

Algorithm 3 Empirical interpolation algorithm

```
function EMPIRICALINTERPOLATION($\{\Xi_q\}_{q=1}^Q$)
    for $q = 1, \dots, Q$ do
        if $q$ is 1 then
            $r \leftarrow \Xi_q$
        else
            $\theta \leftarrow B^{-1}\left[\Xi_q(x_1^{\text{eim}}), \dots, [\Xi_q(x_{q-1}^{\text{eim}})\right]$
            $r \leftarrow \Xi_q - \theta \cdot X$
        end if
        $x_q^{\text{eim}} \leftarrow \arg\max_{x \in \Omega} \|r(x)\|_\infty$
        $X_q \leftarrow r / r(x_q^{\text{eim}})$
        for $q' = 1, \dots, Q$ do
            $B_{q,q'} \leftarrow X_q(x_{q'}^{\text{eim}})$        ▷ $B$ is lower-triangular with unit diagonal
        end for
    end for
    return $\{X_q\}_{q=1}^Q, \{x_q^{\text{eim}}\}_{q=1}^Q, B$ ▷ Interpolation functions, points, and matrix.
end function
```

2.6 Registration methods

Reduced basis approaches build on very well understood and established linear approximation methods and can provide provable error bounds, something that sets them apart from other ROMs which are of a more black-box nature. However, when looking beyond elliptic and parabolic PDEs, one encounters the problem that the n-width of the solution manifold of (even very simple) transport and hyperbolic problems decays as slowly as $\mathcal{O}(n^{-1/2})$ [56, 69, 101]. Linear RB methods are fundamentally not suited to tackle these problems.

Limitations of reduced basis methods

Reduced basis methods can be interpreted as a special class of spectral methods that do not rely on classes of functions with general approximation qualities (polynomials, Fourier spaces, ...) but instead use functions tailored to the problem at hand. As such, they share the shortcomings of spectral methods when approximating jump discontinuities and moving features. The ansatz $u_{\text{rb}}(\mu, x) = \sum_{i=1}^n \mathrm{u}_i(\mu)\zeta_i(x)$ relies on a separation of variables that is not readily applicable to solutions of the form $u(\mu, x) = u_0(x - \mu)$.

Example 2.3 (Examples for fast and slow n-width decay)**.** *Consider two PPDE problems in $\Omega = [0,1]^2 \subset \mathbb{R}^2$ with homogeneous Dirichlet boundary conditions.*

First,

$$\nabla \cdot (K(\mu, x)\nabla u(x)) = 1 \quad \forall x \in \Omega, \tag{2.35}$$

where $x \mapsto K(\mu, x) \in \mathbb{R}_{>0}$ are piece-wise constant functions:

$$K(\mu, \cdot) = \mu_1 \mathbb{1}_{[0,1/2]\times[0,1/2]} + \mu_2 \mathbb{1}_{(1/2,1]\times[0,1/2]} + \mu_3 \mathbb{1}_{(1/2,1]\times(1/2,1]} + \mu_4 \mathbb{1}_{[0,1/2]\times(1/2,1]} \tag{2.36}$$

and $\mu \in [\mu_{\min}, \mu_{\max}]^4$ with $\mu_{\max} = 1/\mu_{\min} = 50$. Second,

$$\Delta u(x) = f(\mu, x) \quad \forall x \in \Omega, \tag{2.37}$$

where $f(\mu,\cdot) = \mathcal{N}(\mu, \mathrm{var})$ *is a narrow Gaussian centered at* $\mu \in [\mu_{\min}, \mu_{\max}]^2$ *with* $\mu_{\max} = -\mu_{\min} = 7/20$*.*

We plot the eigenvalues of $\mathbb{C}^u$ for these two cases in Figure 2.2.

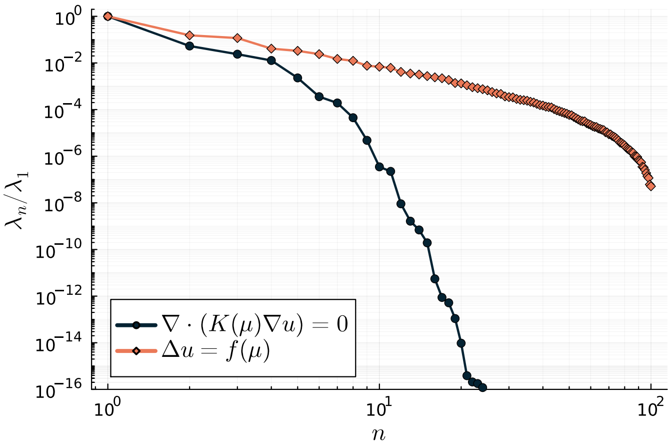

Figure 2.2: Eigenvalues of $\mathbb{C}^u$ for the two cases described in Example 2.3.

The first problem in Example 2.3, where K varies with the parameter value, exhibits fast n-width decay. The eigenvalue energy $\mathcal{E}(n)$ is large already for very small n. For example, $1 - \mathcal{E}(4) < 3 \times 10^{-3}$ and $1 - \mathcal{E}(10) < 3 \times 10^{-7}$. The problem is also parameter-separable and uniformly elliptic (with constant $\mu_{\min}$), so it is extremely favorable to reduction. The second problem, even though it is also elliptic, shows slow n-width decay ($1 - \mathcal{E}(3) > 10^{-1}$ and $1 - \mathcal{E}(10) > 2 \times 10^{-2}$).

Parameter-dependent mappings

One common strategy to overcome the slow n-width decay it is to find a suitable, parameter dependent mapping Φ_μ such that the manifold of mapped solutions

$$\Phi_\mu(\mathcal{M}) := \{u(\mu) \circ \Phi_\mu^{-1} : u(\mu) \in \mathcal{M}\} \tag{2.38}$$

has a much smaller n-width [112, 128, 97, 35, 142, 72]. This approach we refer to as *registration methods.*

Example 2.4. *In the simple one-dimensional case* $\mathcal{M} = \{x \mapsto u_0(x-\mu) : \mu \in \mathbb{R}\}$*, the n-width decays only as* $n^{-1/2}$*, yet the mapped solution manifold* $\Phi_\mu : x \mapsto x - \mu$ *consists of the single element* u_0*.*

As obvious as the previous example is, registration problems are not without challenges. The function $\mu \mapsto \Phi_\mu$ has to be evaluated online, therefore has to be computationally cheap, and return a bijection. Solving Definition 1.1 with the ansatz

$$u_{\mathrm{trb}}(\mu, x) := \sum_{i=1}^{n} \mathrm{u}_i(\mu)\phi_i \circ \Phi_\mu(x) \tag{2.39}$$

for a reduced basis $\{\phi_i\}_{i=1}^n$ built for $\Phi_\mu(\mathcal{M})$ may require taking x-derivatives of Φ_μ, which therefore have to exist and be bounded for numerical stability. Besides these issues of regularity, a good measure for the registration performance of the mapping has to be selected.

Example 2.5. *If, for example, one wants to align two features $f_{x.}$ centered around x_1 and x_2, then a gradient-based minimization of $\|f_{x_1} - f_{x_2} \circ \Phi\|_{L^2}$ may very well fail if x_1 and x_2 are far enough apart such that the support of f_{x_1} and $f_{x_2} \circ \Phi$ does not overlap at the starting point of the optimization.*

Example 2.6. *Consider the set $\mathcal{M} := \{u(\mu) = \mathcal{N}(\mu, \mathrm{var}) : \mu \in [0,1]\}$ of shifted Gaussian distributions on $\mathbb{R}$. One has*

$$\|u(\mu_j) - u(\mu_i)\|_{L^2} = \left(\frac{1}{\sqrt{\pi \mathrm{var}}} - \frac{1}{\sqrt{\pi \mathrm{var}}} \exp\left(-\frac{(\mu_i - \mu_j)^2}{4\mathrm{var}}\right)\right)^{1/2}. \tag{2.40}$$

For $\mathrm{var} \ll 1$, gradients with respect to μ_i of this quantity are essentially zero once $|\mu_i - \mu_j| > 3\sqrt{2\mathrm{var}}$ and the gradient takes extreme values $\propto \mathrm{var}^{-3/4}$ when $\mu_i \approx \mu_j$.

Remark 2.12. *Reduced order modelling applications are only one (fringe) example of registration problems, which are common in imaging sciences, e.g. with medical applications [60]. Broadly speaking, these methods combine* fidelity *or* proximity *terms* Prox(Φ) *and regularization terms* Reg(Φ)*, typically weighted by a hyperparameter, to find an optimal mapping. The result is a minimization of the form*

$$\min_{\Phi} \left(\mathrm{Prox}(\Phi) + \alpha \mathrm{Reg}(\Phi)\right) \tag{2.41}$$

for $\alpha > 0$, coupled with constraints to enforce invertibility and regularity of Φ.

One example, the large deformation diffeomorphic metric mapping *utilizes deformations Φ that are the flow of vector fields from a reproducing kernel Hilbert space and the corresponding norm serves as regularization [134].*

Another example in reduced order modelling that is an application of registration methods are problems with parameter-dependent geometry. When the PPDE problem is formulated on a family of domains $\{\Omega_\mu : \mu \in \mathcal{A}\}$, it is not possible to build a single reduced basis for all parameter values, as the domains of the reduced bases differ. Instead, a mapping $\Phi_\mu(\Omega_0) = \Omega_\mu$ can be used to construct a reduced basis in a reference domain Ω_0, which is then used to solve the PPDE problem.

Map-then-discretize and discretize-then-map

For the sake of exposition, consider the second problem from Example 2.3 in weak form. We are interested in finding $u : \mathcal{A} \to H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u(\mu, x) \cdot \nabla v(x) \mathrm{d}x = \int_\Omega f(\mu, x) v(x) \mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{2.42}$$

Furthermore, assume we have found a family of invertible mappings $\Phi_\mu : \Omega \to \Omega$ such that we expect $\{u(\mu) \circ \Phi_\mu^{-1} : \mu \in \mathcal{A}\}$ to be compressible using a reduced basis approach.

If we let $y = \Phi(x)$, then a Laplace operator in weak form after the registration process can be expressed as

$$\int_\Omega \nabla(\phi_j \circ \Phi_\mu) \cdot \nabla(\phi_j \circ \Phi_\mu) \, \mathrm{d}x = \int_{\Phi_\mu(\Omega)} \nabla \phi_j \cdot [D\Phi_\mu^{-1}]^{-1} [D\Phi_\mu^{-1}]^{-T} \nabla \phi_j |\det D\Phi_\mu^{-1}| \, \mathrm{d}y. \tag{2.43}$$

In the case where the domain $\Omega = \Omega_\mu$ is parameter dependent, the left-hand side of Equation (2.43) will be posed on Ω_μ while the right-hand side will be on Ω_0. This approach is called *map-then-discretize* (MtD) in [130] and has been used in, for example [12, 114].

The weak form in the reference frame is significantly more complex than the original one though the (μ-dependent) matrix $K_\mu := [D\Phi_\mu^{-1}]^{-1}[D\Phi_\mu^{-1}]^{-T} \det D\Phi_\mu^{-1}$. Depending on the complexity of Φ_μ, this means that, for example, a higher order quadrature rule should be used during the assembly step. In reduced order modelling approaches, any method that requires modifications of the high fidelity solver has to be handled with care. In realistic cases, these codes are highly optimized and might not be easily accessible.

If the map Φ_μ is represented by a finite element function itself (as is the case in the example in Chapter 6), it is pointed out in [130] that it is beneficial if the meshes that Φ_μ and $u(\mu)$ are defined on are conforming. If they are not, discontinuities of $D\Phi_\mu$, and therefore K_μ, will drastically decrease the performance of the numerical quadrature.

An alternative method is to still solve the original problem, i.e. the unmodified weak form, and instead move the mesh. As long as the high-fidelity method can handle unstructured meshes, it can be used as-is to compute this *discretize-then-map* (DtM) formulation. Furthermore, the mapping only needs to be evaluated at the mesh nodes, and no derivatives have to be taken. Note the method still requires constraints on $\det D\Phi_\mu$ in order to not produce mesh cells that are very anisotropic or even inverted.

Chapter 3

Optimal transport

Optimal transport (OT) theory provides a notion of discrepancy between probability measures $\rho, \sigma \in \mathcal{P}(\Omega_1) \times \mathcal{P}(\Omega_2)$. In particular, a cost is modelled through a function $c : \Omega_1 \times \Omega_2 \to \mathbb{R}$ where $c(x, y)$ gives the cost of moving a differential unit of mass from x to y. The larger the cost to move the total mass, the more different the two probability densities are to one another. This defines a distance on the space of probability densities with a number of appealing properties.

With our application in mind, we will make the following simplifying assumptions: Ω_1 and Ω_2 are bounded subsets of $\mathbb{R}^d$. In most cases, we will consider the case $\Omega_1 = \Omega_2 = \Omega$. We remark that optimal transport theory has been developed for much more general spaces and (lower semi-continuous) cost functions.

Furthermore, we will focus on the choice $c(x, y) = \frac{1}{2}|x - y|^2$ and where the measures ρ, σ admit a density with respect to the Lebesgue measure, i.e. they are absolutely continuous. This will be explicitly started when required.

The research on OT theory and its applications in physics, economics, imaging sciences, et cetera is extensive. Several excellent textbooks and lecture notes on the topic are available. Villani's works can be called standard references in the field [140, 139]. A stronger emphasis on the relation to partial differential equations is made in [115, 57] and computational aspects are thoroughly treated in [106, 138].

3.1 Short history of the optimal transport problem

The inception of optimal transport is usually attributed to French geometer Gaspard Monge[1] in the year 1781.

The Monge problem

Motivated by the question of how to optimally extract construction materials, he asked what is the optimal way to move, say, a pile of sand from one configuration (*déblai*) into another (*remblai*). The optimization is an assignment problem, without loss of generality we can assume that the sand piles have unit mass and are modelled by probability densities.

[1] Monge was also a Minister of the Marine, founder of the École Polytechnique, close friend of Napoleon Bonaparte, and once won a race up a pyramid ([139], Chapter 3, [65], Appendices I)

Definition 3.1 (Monge problem)**.** *Given two probability measures $\rho, \sigma \in \mathcal{P}(\Omega)$, find the optimal transport* map $T : \Omega \to \Omega$ *solving*

$$\inf_{T:\sigma=T_\sharp\rho} \int c(x, T(x))\mathrm{d}\rho. \tag{3.1}$$

Here, $T_\sharp\rho$ denotes the push-forward measure of ρ under T, defined as

Definition 3.2 (Push-forward)**.** *The push-forward of ρ under $T : \Omega \to \Omega$, denoted by $T_\sharp\rho$, is defined by $(T_\sharp\rho)[\Omega'] = \rho[T^{-1}(\Omega')]$ for all measurable $\Omega' \subset \Omega$. If both ρ and σ are absolutely continuous with densities denoted again ρ, σ, and T is a C^1 diffeomorphism, then $\sigma = T_\sharp\rho$ is equivalent to*

$$\rho = (\sigma \circ T)\,|\det DT|. \tag{3.2}$$

In Monge's original work he assumed $c(x, y) = |x - y|$, a natural choice but ambitious to treat as this cost function is not strictly convex.

The optimization problem Equation (3.1) is challenging. The functional and the constraint are non-linear, the latter non-local. If we assume that T is a smooth diffeomorphism, it is equivalent to the fully non-linear partial differential equation $\rho(x) = (\sigma \circ T(x))\,|\det DT(x)|$, which has to hold ρ-almost everywhere. It is not clear if a solution to this problem exists. In fact:

Example 3.1. *Let $\rho = \delta_x$ and σ any measure that is not of the form δ_y for some $y \in \Omega$. As $T_\sharp\delta_x = \delta_{T(x)}$, no transport map exists.*

The Kantorovich problem

In 1942, Leonid Kantorovich introduced a similar transport problem that can be seen as a relaxation of Monge's. He himself noted this connection in 1948 [76].

Definition 3.3 (Kantorovich problem)**.** *Given two probability measures $\rho, \sigma \in \mathcal{P}$, find the optimal transport* plan $\pi \in \mathcal{P}(\Omega \times \Omega)$ *solving*

$$\inf_{\pi\in\Pi(\rho,\sigma)} \int_{\Omega\times\Omega} c(x, y)\mathrm{d}\pi(x, y), \tag{3.3}$$

where $\Pi(\rho, \sigma)$ is the set of admissible transport plans with fixed marginals:

$$\Pi(\rho, \sigma) := \{\pi \in \mathcal{P}(\Omega \times \Omega) : \pi(\cdot, \Omega) = \rho \textit{ and } \pi(\Omega, \cdot) = \sigma\}. \tag{3.4}$$

We will refer to c as the *cost function* and to the value $\int c\mathrm{d}\pi$ as the *(total) transport cost* corresponding to the plan π.

Remark 3.1. *The set $\Pi(\rho, \sigma)$ is non-empty, as it always contains the product measure $\rho \otimes \sigma$. Existence of a solution to Equation* (3.3) *can be shown for very general cost functions (lower semi-continuous and bounded from below), see for example [115], Theorem 1.5.*

For $\Omega', \Omega'' \subset \Omega$, the value of $\pi(\Omega', \Omega'')$ is the amount of mass that is transferred from Ω' to Ω''. Kantorovich's formulation therefore allows mass splitting, something Monge ruled out by requiring the existence of a transport map T, which is not multivalued. Kantorovich's formulation is a generalization of Monge's problem in the following sense:

Remark 3.2. *If a transport plan π is of the form $(\mathrm{id}, T)_\sharp\rho$, i.e. supported on the graph $(x, T(x))$, then Equation* (3.3) *takes the form of Equation* (3.1).

Duality

The Kantorovich problem consists of the minimization of a linear functional over a convex set, $\mathcal{P}(\Omega \times \Omega)$. As such, it admits a dual problem:

Theorem 3.1 (Kantorovich duality ([140], Theorem 1.3))**.** *The dual problem of Equation* (3.3) *is given by*

$$\sup_{\psi_\rho, \psi_\sigma \in \mathcal{C}(\Omega)} \left\{ \int_\Omega \psi_\rho \mathrm{d}\rho + \int_\Omega \psi_\sigma \mathrm{d}\sigma \ : \ \psi_\rho(x) + \psi_\sigma(y) \leq c(x, y) \right\}, \tag{3.5}$$

where $c : \Omega \times \Omega \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ *is a lower semi-continuous cost function. The supremum in Equation* (3.5) *and the infimum in Equation* (3.3) *are equal, and both are attained.*

We will refer to the functions ψ_ρ, ψ_σ as *transport potentials.* Those potentials that realize the maximum in Equation (3.5) we will refer to as *optimal transport potentials.*

Remark 3.3. *While it is somewhat natural to consider transport potentials in* $\mathcal{C}_b(\Omega)$*, as continuous functions are in duality with measures, it is not obvious a priori that optimal transport maps are continuous instead of, for example, elements of* $L^1(\rho)$ *and* $L^1(\sigma)$*. As we will see, the potentials in fact inherit their continuity from the cost function.*

This duality has an intuitive interpretation, cited here after [140], who credits Caffarelli: The primal problem Equation (3.3) corresponds to a centralized distribution of goods from suppliers ρ to consumers σ.

The goal is to assign the goods in a way such that the transport cost is minimal and at the same time the entire supply is used and the demand is met.

The dual problem Equation (3.5) solves this by delegating it to a logistics company that charges a fee ψ_ρ to pick up the goods at the suppliers and another fee ψ_σ to drop them off at the consumers (both ψ_ρ and ψ_σ can also be negative in some places). The logistics company wants to maximize its profit but operates under the constraint that it cannot charge more than what the cost would be if the goods were transported by the central actor from the primal problem: $\psi_\rho(x) + \psi_\sigma(y) \leq c(x, y)$.

Transport induced by maps

What remains is the question when the Monge and Kantorovich formulation coincide, i.e. when the transport plan is in fact concentrated on a graph and there exists an optimal transport map.

Theorem 3.2 (Brenier's theorem)**.** *If* $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega)$*, then the unique solution to Equation* (3.3) *with quadratic cost is concentrated on the graph* $(x, T(x))$ *of a transport map* T*. In particular,*

$$\inf_{\pi \in \Pi(\rho,\sigma)} \int_{\Omega \times \Omega} |x - y|^2 \mathrm{d}\pi(x, y) = \inf_{S : \sigma = S_\sharp \rho} \int_\Omega |S(x) - x|^2 \mathrm{d}\rho(x). \tag{3.6}$$

The minimum is attained by an optimal T, which is the gradient of a Lipschitz continuous convex function:

$$T(x) = \nabla\varphi = x - \nabla\psi(x). \tag{3.7}$$

The function ψ is the optimal transport potential of Equation (3.5)*, denoted ψ_ρ therein.*

Remark 3.4. *The result holds for more general cost functions satisfying a* twist condition*, i.e. $x-$differentiability and injectivity of $y \mapsto \nabla_x c(x,y)$, see [115], Section 1.3.*

3.2 Dual problem

We do not repeat here the proof of Kantorovich's duality result, referring to e.g. [115], Section 1.6. We do, however, repeat a formal calculation performed therein that provides some context for the form of the dual problem. Note that

$$\sup_{\psi_\rho,\psi_\sigma\in\mathcal{C}} \left(\int_\Omega \psi_\rho \mathrm{d}\rho + \int_\Omega \psi_\sigma \mathrm{d}\sigma - \int_{\Omega\times\Omega} (\psi_\rho + \psi_\sigma)\mathrm{d}\pi \right) = \begin{cases} 0 & \text{if } \pi \in \Pi(\rho,\sigma) \\ +\infty & \text{else.} \end{cases} \tag{3.8}$$

We now add the marginal constraints into the objective function to obtain the following formulation, equivalent to Equation (3.3):

$$\inf_{\pi\in\mathcal{P}(\Omega\times\Omega)} \left(\int_{\Omega\times\Omega} c\,\mathrm{d}\pi + \sup_{\psi_\rho,\psi_\sigma\in\mathcal{C}} \left(\int_\Omega \psi_\rho \mathrm{d}\rho + \int_\Omega \psi_\sigma \mathrm{d}\sigma - \int_{\Omega\times\Omega} (\psi_\rho \oplus \psi_\sigma)\mathrm{d}\pi \right) \right). \tag{3.9}$$

If it is possible to exchange the sup and inf in this term - and the duality proof is concerned with just that - then we arrive at

$$\sup_{\psi_\rho,\psi_\sigma} \left(\int_\Omega \psi_\rho \mathrm{d}\rho + \inf_\pi \left(\int_{\Omega\times\Omega} (c - (\psi_\rho \oplus \psi_\sigma))\mathrm{d}\pi \right) \right). \tag{3.10}$$

Since

$$\inf_\pi \int_{\Omega\times\Omega} (c - (\psi_\rho \oplus \psi_\sigma))\mathrm{d}\pi = \begin{cases} 0 & \text{if } c - (\psi_\rho \oplus \psi_\sigma) \le 0 \\ -\infty & \text{else,} \end{cases} \tag{3.11}$$

we obtain Equation (3.5).

The dual problem is appealing as the constraint is linear and so is the space $\mathcal{C}(\Omega)$.

c-transform

Assume that we are given a candidate potential ψ_ρ. In order to maximize the objective of Equation (3.5), it is natural to choose as ψ_σ the largest possible function (recall that σ is non-negative) that does not violate the constraint:

Definition 3.4 (c-transform)**.** *The c-transform of a function $\psi : \Omega \to \mathbb{R} \cup \{+\infty\}$ is given by*

$$\psi^c(y) := \inf_{x\in\Omega} \left(c(x,y) - \psi(x)\right). \tag{3.12}$$

Going further, we can replace ψ_ρ by $\psi_\sigma^c = \psi_\rho^{cc}$. From the definition of the c-transform, this will only increase the value of the integrals in the dual problem. We could go on, but $\psi_\rho^{ccc} = \psi_\rho^c$ as, for any ψ_1, ψ_2, we have $\psi_1^{cc} \geq \psi_1$ and $\psi_1 \geq \psi_2 \Rightarrow \psi_1^c \leq \psi_2^c$.

We say a function ψ_1 is *c-concave* if there exists ψ_2 such that $\psi_1 = \psi_2^c$. Furthermore, ψ and ψ^c are called *c-conjugate*. In the following, we will often drop the subscript on ψ once it is established what is the initial and what is the target measure.

With the notion of c-transform, we have a strategy at hand to maximize the objective of the dual problem:

Proposition 3.1 ([115], Proposition 1.11)**.** *Assume c is continuous and bounded. The dual problem Equation* (3.5) *admits a solution of the form*

$$\max_{\psi_\rho \ c\text{-concave}} \int \psi_\rho \mathrm{d}\rho + \int \psi_\rho^c \mathrm{d}\sigma. \tag{3.13}$$

The proof of Proposition 3.1 is omitted here for brevity. The strategy is as follows: start with a maximizing sequence $(\psi_\rho^{(n)}, \psi_\sigma^{(n)})$. Applying the c-transform to the sequence elements evidently only improves it. One can then use the definition of the c-transform to show that the modified sequence shares its modulus of continuity with the cost function and derive uniform bounds. Therefore, the sequence is equicontinuous and equibounded and one can apply the theorem of Ascoli-Arzelà to obtain a uniformly convergent subsequence.

Remark 3.5. *The fact that the optimal transport potential ψ_ρ shares its modulus of continuity with c implies that it is differentiable almost everywhere by Rademacher's Theorem ([70], Theorem 4.2.3). By the assumptions in Theorem 3.2, almost everywhere with respect to the Lebesgue measure is ρ-almost everywhere, so $\nabla\psi_\rho$ is well-defined.*

Convexity and the quadratic cost function

When dealing with the quadratic cost, we can relate the c-transform to the known notion of the Legrendre transform of convex functions. Indeed, convexity takes the place of c-concavity and leads to the result of Brenier's Theorem 3.2:

Proposition 3.2. *In the case of $c(x,y) = \frac{1}{2}|x-y|^2$, it holds that $\frac{|y|^2}{2} - \psi^c(y)$ is the Legendre transform of $\varphi(x) := \frac{|x|^2}{2} - \psi(x)$, defined by*

$$\varphi^*(y) := \sup_{x\in\mathbb{R}^d}(x \cdot y - \varphi(x)). \tag{3.14}$$

Proof.

$$\frac{|y|^2}{2} - \psi^c(y) = \frac{|y|^2}{2} - \inf_{x\in\Omega}\left(\frac{1}{2}|x-y|^2 - \psi(x)\right) = \sup_{x\in\Omega}\left(x\cdot y - \frac{|x|^2}{2} + \psi(x)\right). \tag{3.15}$$

□

Remark 3.6. *The factor $\frac{1}{2}$ leads to the convenient form of $\nabla\varphi = \mathrm{id} - \nabla\psi$. However, including this factor is not standard, hence we will refer to $|x-y|^2$ when we speak of the quadratic cost case. Naturally this factor can be absorbed in the potentials ψ if needed. However, when we write ψ and ψ^c, we will always do so with respect to the cost $c(x,y) = \frac{1}{2}|x-y|^2$, therefore*

$$W_2(\rho,\sigma)^2 = 2\int \psi_\rho \mathrm{d}\rho + 2\int \psi_\sigma \mathrm{d}\sigma, \tag{3.16}$$

where ψ_ρ, ψ_σ are the optimal transport potentials, i.e. solutions of the dual problem Equation (3.5)*.*

In summary: φ, φ^* are the convex functions whose gradients give the transport maps, while $\psi(x) = \frac{|x|^2}{2} - \varphi(x)$ and $\psi^c(y) = \frac{|y|^2}{2} - \varphi^*(y)$.

Remark 3.7. *Note that the Legendre transform is typically defined on the entirety of $\mathbb{R}^d$. Likewise, the value of the transport potentials is only determined ρ- (respectively σ-) almost everywhere by the optimal transport dual problem. Through the c-transform, one can extend the domain of the potentials to Ω or even $\mathbb{R}^d$.*

Remark 3.8. *When Brenier's theorem is applicable, this result implies that we can invert the mapping $T : x \mapsto \nabla\varphi(x)$ by evaluating the c-transform. As we will see, this can be an interesting option in numerical applications.*

Proposition 3.3. *Assume $\varphi : \mathbb{R}^d \to \mathbb{R}$ is strictly convex, differentiable, and increases faster than any linear function as $|x| \to +\infty$. Then, $(\nabla\varphi)^{-1} = \nabla\varphi^*$.*

Proof. By the assumptions, there exists for every y a unique $x^*(y) \in \mathbb{R}^d$ where the minimum is attained, characterized by the condition $y = \nabla\varphi(x^*(y))$. At the same time, $\nabla\varphi^*(y) = x^*(y) - (\nabla\varphi(x^*(y)) - y) \cdot \nabla x^*(y) = x^*(y)$ and thus $\nabla\varphi(x) = y$ if and only if $\nabla\varphi^*(y) = x$. □

Sufficient criteria

After establishing that optimal transport maps of densities necessarily have gradient structure, it is a natural question if this relation holds both ways.

Theorem 3.3 (Sufficient criterion for optimal maps ([115], Theorem 1.48))**.** *Suppose $\rho \in \mathcal{P}_{\mathrm{ac}}$ and $\varphi : \Omega \to \mathbb{R} \cap \{+\infty\}$ is a convex function, differentiable almost everywhere. Then, $\nabla\varphi$ is the optimal transport map between ρ and $\nabla\varphi_\sharp\rho$.*

Proof. Recall from the definition of the Legendre transform that $\varphi(x) + \varphi^*(y) \geq x \cdot y$ for all $x, y \in \mathbb{R}^d$. Furthermore, equality only holds if $\nabla\varphi(x) = y$. Let $\nabla\varphi_\sharp\rho =: \sigma$. For any admissible transport plan between ρ and σ, we have

$$\frac{1}{2}\int |x-y|^2 \mathrm{d}\pi(x,y) = \frac{1}{2}\int |x|^2 \mathrm{d}\rho + \frac{1}{2}\int |x|^2 \mathrm{d}\sigma - \int x \cdot y \mathrm{d}\pi(x,y). \tag{3.17}$$

The first two terms do not depend on the plan. For the last term, we can use the bound

$$\begin{aligned}
\int x \cdot y \mathrm{d}\pi(x,y) \leq \int (\varphi(x) + \varphi^*(y)) \mathrm{d}\pi(x,y) \qquad & (3.18)\\
= \int \varphi(x)\mathrm{d}\rho + \int \varphi^*(y)\mathrm{d}\sigma(y) \qquad & (3.19)\\
= \int (\varphi + \varphi^* \circ \nabla\varphi)\mathrm{d}\rho. \qquad & (3.20)
\end{aligned}$$

Equality is attained by the plan induced by the map $x \mapsto \nabla\varphi(x)$, hence the latter is optimal. □

Remark 3.9. *When ρ and σ do not allow a transport map, the condition that $T = \nabla\varphi$ has a counterpart: When π is the optimal transport plan between ρ and σ, then π is supported in the subdifferential of a proper lower semi-continuous convex function ([140], Theorem 2.29). If furthermore the total cost is finite for the trial plan $\rho \otimes \sigma$, this condition is also sufficient (see the discussion after Proposition 2.24 therein).*

Furthermore, by Rockafellar's Theorem ([140], Theorem 2.27), this implies that the support of π is cyclically monotone*. This relation is in fact if and only if.*

Definition 3.5 (Cyclical monotonicity)**.** *A set $\Gamma \subset \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone if for all $m \in \mathbb{N}$ and for all $(x_1, y_1), \ldots, (x_m, y_m) \in \Gamma$*

$$\sum_{i=1}^{m} |x_i - y_i|^2 \leq \sum_{i=1}^{m} |x_i - y_{\mathrm{perm}(i)}|^2 \tag{3.21}$$

for any permutation perm.

To interpret this condition, assume that we have a trial transport plan π'. Looking at the pair of points $(x_1, y_1) \in \mathrm{supp}(\pi')$, we can lower the total cost by instead linking x_1 and y_i, where $|x_1 - y_i|$ is strictly smaller than $|x_1 - y_1|$. Say this lowers the total cost by δ. Naturally, the marginal constraints are now violated: there is an excess of mass at y_i and a deficit at y_1. Hence, we have to find another point, which can without loss of generality be called x_2, that is linked with y_i in the trial plan. We then link x_2 with another point y_j and so on. Eventually we close this cycle to restore the marginal constraints when we link x_m to y_1. If we increase the cost only by $\delta' < \delta$ during this correction process, we have found a permutation perm that improves the trial plan:

$$\sum_{i=1}^{m} |x_i - y_i|^2 > \sum_{i=1}^{m} |x_i - y_{\mathrm{perm}(i)}|^2. \tag{3.22}$$

3.3 Monge-Ampère equation

From Theorem 3.3 (Sufficient criterion for the optimality of a transport map), one can derive a Euler-Lagrange equation for the optimal transport problem. Assume that $\varphi \in \mathcal{C}^2(\Omega)$ is strictly convex. Then, if $\nabla\varphi$ pushes $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ forward to $\sigma \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, it necessarily holds that (c.f. Equation (3.2)):

$$\rho(x) = \sigma(\nabla\varphi(x)) \det D^2\varphi(x) \quad \forall x \in \Omega. \tag{3.23}$$

If we furthermore assume that σ is strictly positive, then

$$\det D^2\varphi(x) = \frac{\rho(x)}{\sigma \circ \nabla\varphi(x)} \quad \forall x \in \Omega. \tag{3.24}$$

This is known as a *Monge-Ampére* type equation, which take the form

$$\det D^2\varphi(x) = f(x, \varphi, \nabla\varphi) \tag{3.25}$$

for some function f. For the purposes of this work, we are only interested in the specific instance given in Equation (3.24), and will refer it as *the* Monge-Ampére equation.

Properties

Equation (3.24) comes without a boundary condition, it only requires that $\nabla\varphi$ maps the support of ρ to that of σ. To be more precise, let Ω_1 and Ω_2 be two open, bounded subsets of $\mathbb{R}^d$ such that $\operatorname{supp}(\rho) = \overline{\Omega}_1$, $\operatorname{supp}(\sigma) = \overline{\Omega}_2$, and both $\partial\Omega_1$ and $\partial\Omega_2$ are of Lebesgue measure zero. Furthermore, assume that ρ and σ are bounded away from both zero and infinity on their supports. Then, Equation (3.24) together with the condition

$$\nabla\varphi(\Omega_1) = \Omega_2, \tag{3.26}$$

is called a *second boundary value problem* for the Monge-Ampére equation. If $\nabla\varphi$ is continuous with continuous inverse, then this condition is equivalent to $\nabla\varphi(\partial\Omega_1) = \partial\Omega_2$ ([115], Section 1.7.6).

The Monge-Ampére equation is a so-called *fully non-linear elliptic equation* (c.f. [140], Definition 4.2). The ellipticity of the equation in this sense is a consequence of the monotonicity of the determinant in the following sense: If $A - B$ is a negative semi-definite matrix (write $A \leq B$), then $\det A \leq \det B$. However, this monotonicity is very degenerate.

Even in the simplest case where the transported densities are uniform on their respective supports, and $|\operatorname{supp}(\rho)| = |\operatorname{supp}(\sigma)|$, the resulting equation $\det D^2\varphi(x) = 1$ is invariant under a number of transformations including any affine transformation with determinant equal to one. For example, if φ solves $\det D^2\varphi(x) = 1$ on $\mathbb{R}^d$, then so does $(x, y) \mapsto \varphi(\epsilon x, y/\epsilon)$ for any $\epsilon > 0$ (see for example [47], Section 2.3).

Linearization

The ellipticity of Equation (3.24) is clearer in the linearized case or when the transport is (formally) small, see [140], Exercise 4.1. Recall that $\nabla\varphi_\sharp\rho = \sigma$ corresponds to

$$(\sigma \circ \nabla\varphi) \det D^2\varphi = \rho \tag{3.27}$$

in the case where $\nabla\varphi$ is C^1 and strictly convex. Assume for now that ρ is a smooth, strictly positive density that is transported to the density $\sigma^\epsilon = \rho(1 - \epsilon\nu + \mathcal{O}(\epsilon^2))$. If we make the ansatz $\phi(x) = \frac{|x|^2}{2} - \epsilon\psi(x) + \mathcal{O}(\epsilon^2)$, then Equation (3.24) becomes

$$(\rho(x) - \epsilon\nabla\rho(x) \cdot \nabla\psi - \epsilon\nu(x)\rho(x))(1 - \epsilon\Delta\psi(x)) = \rho(x) + \mathcal{O}(\epsilon^2). \tag{3.28}$$

Neglecting the second order-terms, ψ is the solution to the following linear elliptic problem in weak form:

$$\int \nabla\psi \cdot \nabla\chi \, \mathrm{d}\rho = \int \chi\nu \, \mathrm{d}\rho \quad \forall\chi \in \mathcal{C}_0^\infty(\Omega). \tag{3.29}$$

Weak solutions

A rigorous study of the Monge-Ampére equation requires the introduction of weak solutions, since we do not know if φ is twice differentiable, i.e. the left hand side of Equation (3.24) is measure-valued.

In the setting of Theorem 3.2, we only know that φ is differentiable almost everywhere. In this case, one needs to generalize what is $\det D^2\varphi$ in the smooth case. One way to do so is the following: Define the *Hessian measure*

$$\det_H D^2\varphi : \det_H D^2\varphi[\Omega'] = \left|\bigcup_{x\in\Omega'} \partial\varphi(x)\right| \quad \text{for all measurable sets } \Omega' \subset \mathbb{R}^d. \tag{3.30}$$

If $\det_H D^2\varphi$ is absolutely continuous and Equation (3.24) holds almost everywhere, then φ is an *Alexandrov solution* of the Monge-Ampére equation.

The weakest of the commonly used notions is that if $\nabla\varphi_\sharp\rho = \sigma$, one calls φ a *Brenier solution* of Equation (3.24).

The theory of regularity of optimal transport maps is thus closely linked to the study of the regularity of the Monge-Ampére equation. This is a very involved topic and a discussion that does it justice is beyond the scope of this thesis. An introduction is given in the optimal transport references [115] (Section 1.7.6) and [140] (Chapter 4). For a more detailed overview, we refer to [48, 47] as well as [135] and [139], Chapter 12.

We will require some of the results from this theory when employing the optimal transport maps in Chapter 6.

3.4 Regularity

We already know that the transport potential ψ from ρ to σ introduced in Brenier's Theorem 3.2 is differentiable almost everywhere as long as ρ is absolutely continuous.

One-dimensional case

In one spatial dimension, computing the optimal transport distance reduces to a sorting problem. The gradients of functions in this case are non-decreasing functions and the Monge-Ampére equation for the transport map T from a density ρ to σ reads

$$\rho = (\sigma \circ T)\partial_x T(x). \tag{3.31}$$

Introducing the cumulative distribution function valued

$$\operatorname{cdf}(\rho)(x) := \int_{-\infty}^{x} \mathrm{d}\rho, \tag{3.32}$$

integration of Equation (3.31) gives

$$\operatorname{cdf}(\rho)(x) = \int_{-\infty}^{x} \mathrm{d}\rho(y) \int_{-\infty}^{x} \partial_y T(y)\, \mathrm{d}(\sigma \circ T(y)) = \operatorname{cdf}(\sigma) \circ T(x). \tag{3.33}$$

The cumulative distribution function might not be invertible, but it is always non-decreasing, so we can define its pseudo-inverse as

$$\operatorname{cdf}[\rho]^{[-1]}(t) : \inf\{x \in \mathbb{R} : \operatorname{cdf}[\rho](x) \geq t\}, \tag{3.34}$$

which allows us to obtain the optimal transport map through the explicit form

$$T_{\rho\to\sigma} = \operatorname{cdf}(\sigma)^{[-1]} \circ \operatorname{cdf}(\rho). \tag{3.35}$$

Hence, if $\Omega \subset \mathbb{R}$, it follows that $T_{\rho\to\sigma} = \operatorname{cdf}(\sigma)^{[-1]} \circ \operatorname{cdf}(\rho)$ for two measures ρ, σ. The cumulative distribution function of a measure ρ might be discontinuous (if ρ has atoms) or not strictly increasing (if ρ does not have full support in the domain). If both ρ and σ admit continuous densities that are bounded away from zero, then we can write $T_{\rho\to\sigma} = \operatorname{cdf}(\sigma)^{-1} \circ \operatorname{cdf}(\rho)$ and this map is a C^1 diffeomorphism.

Counter-examples

Moving to $d > 1$, let us recall one of the most famous counterexamples to regularity of the optimal map even if the involved measures are very smooth and bounded from above and below. It is attributed to Caffarelli [34]. We recall the version given in [139], Theorem 12.3.

Example 3.2. *Consider $\rho \propto \mathbb{1}_{B(0,1)}$ constant on the unit ball in $\mathbb{R}^2$ centered at zero and $\sigma \propto \frac{1}{2}\mathbb{1}_{B(\hat{e}_1,1)\cap\{x_1>1\}} + \frac{1}{2}\mathbb{1}_{B(-\hat{e}_1,1)\cap\{x_1<1\}}$ two half balls that are shifted by one unit in positive and negative x_1 direction, respectively. The optimal map in this case cannot be continuous, indeed it is given by $T_{\rho\to\sigma}(x) = \nabla(|x|^2/2 + |x_1|)$.*

The disjoint support of the target measure is an obvious obstacle to the continuity of T. However, as it turns out, the non-convexity of the support of the target measure is already enough.

Let us connect the two halves by a strip of mass of diameter δ and denote the resulting set D_δ. By the continuity of optimality, the optimal transport maps T_δ that map ρ into $\sigma_\delta \propto \mathbb{1}_{D_\delta}$ will converge to $T = \nabla(|x|^2/2 + |x_1|)$ as $\delta \to 0$.

We argue that as soon as δ is small enough, these maps can no longer be continuous. Take a small section S near the top of the ball ($x_2 \approx 1$). Since T splits this set to the top halves of the left and right half-ball, as soon as δ is small enough, T_δ also has to map a large amount of the mass in S to the left and right. But since the image of a connected set (which S is) under a continuous map is connected, this implies that $T_\delta(S)$ connects the top of the two connected half-balls and extends across the bridge connecting them.

Clearly, for small δ, this cannot be optimal. Indeed, it would require to map some points $\{x_V\}$ of S a far distance downwards, while all other points $\{x_H\}$ in S are mapped almost purely horizontally. There must some x_H that lies above some x_V, indeed if the x_2 component of all x_V would be smaller than some $x_2^* < 1$, then the map would be discontinuous for all $x : x_2 > x_2^*$. Therefore, there exists two points x_H, x_V such that

$$\begin{aligned}&(x_H - x_V)\cdot(T_\delta(x_H) - T_\delta(x_V)) < 0\\ &\qquad\Leftrightarrow |x_H - T_\delta(x_H)|^2 + |x_V + T_\delta(x_V)|^2 > |x_H - T_\delta(x_V)|^2 + |x_V - T_\delta(x_H)|^2. \quad (3.36)\end{aligned}$$

However, this contradicts the cyclical monotonicity (c.f. Definition 3.5) of the optimal plan that T_δ induces.

While the non-convexity of the target set might be not extreme enough to induce a discontinuity, it is enough to rule out continuous transport in general.

Theorem 3.4 (Existence of discontinuous transport maps ([49], Theorem 1.1))**.** *If Ω_1 and Ω_2 are bounded, open subsets of $\mathbb{R}^d$ and Ω_2 is not convex, there exist smooth densities $\rho \in \mathcal{P}(\Omega_1), \sigma \in \mathcal{P}(\Omega_2)$, bounded away from zero and bounded, such that the optimal transport map from ρ to σ is not continuous.*

Sufficient conditions

At the same time, one can show that convexity of the support of the target measure is sufficient to obtain regularity of the optimal transport map for the quadratic cost of $\mathbb{R}^d$.

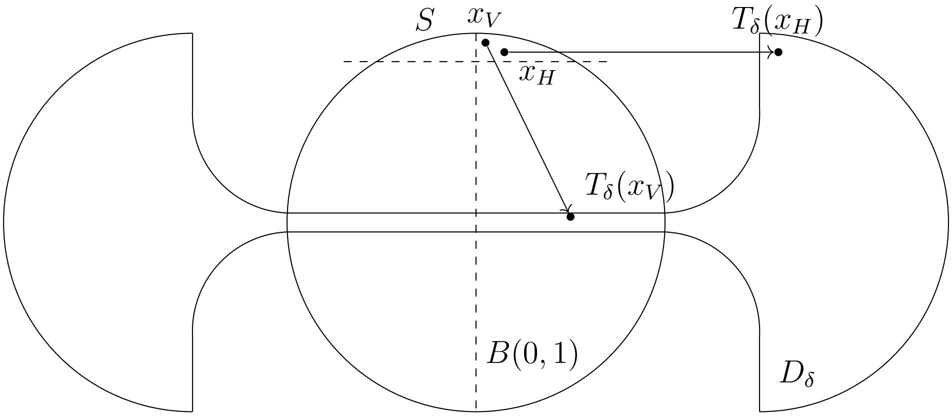

Figure 3.1: Illustration of Example 3.2.

Theorem 3.5 (Caffarelli's regularity theory ([139], Theorem 12.50))**.** *Let $\varphi : \Omega \to \mathbb{R}$ be a convex function such that $\nabla\varphi$ is the optimal transport map from $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega_1)$ to $\sigma \in \mathcal{P}_{\mathrm{ac}}(\Omega_2)$, where Ω_1, Ω_2 are bounded, connected, and open subsets of $\mathbb{R}^d$ and ρ, σ are bounded and bounded from below.*

If Ω_2 is convex, then $\varphi \in \mathcal{C}^{1,\beta}(\Omega_1)$ for some $\beta \in (0,1)$. If furthermore the densities ρ, σ are $\mathcal{C}^{k,\alpha}$ for some $k \in \mathbb{N}_0$ and $\alpha \in (0,1)$, then $\varphi \in \mathcal{C}^{2,\alpha}(\Omega_1)$.

Remark 3.10. *The results in Theorem 3.5 can be tightened further, see Chapter 2 in [47] or Chapter 12 in [139]. In particular, the regularity can be extended to the boundary of the domain under sufficient regularity thereof.*

Remark 3.11. *We saw in Example 3.2 that T was smooth almost everywhere, i.e. the set of singularities was supported on a line. This holds more generally: if the transported densities are smooth, then the transport map is a smooth diffeomorphism between open subsets of Ω_1, Ω_2 and the respective complements of these subsets are of Lebesgue measure zero ([49], Theorem 1.2, see also [67, 61]).*

3.5 Optimal transport metric

Definition 3.6 (Wasserstein distance)**.** *We define the* Wasserstein-p distance *for $p \in [1,\infty)$ between two probability measures $\rho, \sigma \in \mathcal{P}(\Omega)$ by*

$$W_p(\rho,\sigma)^p := \min_{\pi\in\Pi(\rho,\sigma)} \int_{\Omega\times\Omega} |x-y|^p \mathrm{d}\pi(x,y), \tag{3.37}$$

where Π is the set of admissible transport plans as defined in Definition 3.3.

Remark 3.12. *The name Wasserstein distance is named after Leonid Vaseršteĭn, who did however not play a big role in the development of the theory [137]. In any case, the name is established by now and we will make no attempt to change this. Other names include* Kantorovich-Rubenstein distance *or* optimal transport distance. *We will use the terms metric and distance interchangeably.*

Proposition 3.4. *W_p is a metric on $\mathcal{P}(\Omega)$.*

Proof. We will give a proof of this fact only for the case $p = 2$ and when ρ and σ admit a transport map and refer to [115], Section 5.1 for the general case. Positivity and symmetry of W_2 is clear from its definition as is $W_2(\rho,\rho) = 0$ by taking $T = \mathrm{id}$. $W_2(\rho,\sigma) = 0$ implies $T = \mathrm{id}$ ρ-a.e. and thereby $\rho = \sigma$. For the triangle inequality, consider a third, arbitrary density ν. Denote by $T_{\rho\to\nu}$ and $T_{\nu\to\sigma}$ the optimal transport maps between ρ and ν and ν and σ, respectively. Then,

$$W_2(\rho,\sigma)^2 \le \int |T_{\nu\to\sigma} \circ T_{\rho\to\nu}(x) - x|^2 \mathrm{d}\rho \tag{3.38}$$

$$= \|T_{\nu\to\sigma} \circ T_{\rho\to\nu} - \mathrm{id}\|^2_{L^2(\rho)} \tag{3.39}$$

$$\le \|T_{\nu\to\sigma} \circ T_{\rho\to\nu} - T_{\rho\to\nu}\|^2_{L^2(\rho)} + \|T_{\rho\to\nu} - \mathrm{id}\|^2_{L^2(\rho)} \tag{3.40}$$

The first term describes an admissible competitor for the transport from σ to ν and is therefore bounded from above by $W_2(\sigma,\nu)^2$ while the second term equals $W_2(\nu,\rho)^2$. □

Proposition 3.5. *W_2 factors out translations in the following sense: Let $\mathrm{Shift}_a : x \mapsto x - a$ be the shift operator by a vector a and denote by $m_\rho = \int x \mathrm{d}\rho(x)$ the mean of ρ, where ρ is a probability measure on $\mathbb{R}^d$ with finite second moment, as is σ. Then,*

$$W_2(\mathrm{Shift}_{a,\sharp}\rho, \mathrm{Shift}_{a',\sharp}\sigma)^2 = W_2(\rho,\sigma)^2 - 2(a - a') \cdot (m_\rho - m_\sigma) + |m_\rho - m_\sigma|^2. \tag{3.41}$$

In particular, $W_2(\rho,\sigma)^2 = |m_\rho - m_\sigma|^2 + W_2(\rho_o,\sigma_o)^2$ where $a = m_\rho, a' = m_\sigma$, and ρ_o, σ_o are centered at the origin.

Proof.

$$W_2(\mathrm{Shift}_{a,\sharp}\rho, \mathrm{Shift}_{a',\sharp}\sigma)^2 = \min_{\pi\in\Pi(\mathrm{Shift}_{a,\sharp}\rho, \mathrm{Shift}_{a',\sharp}\sigma)} \int |x - y|^2 \mathrm{d}\pi(x,y) \tag{3.42}$$

$$= \min_{\pi\in\Pi(\rho,\sigma)} \int |x + a - y - a'|^2 \mathrm{d}\pi(x,y) \tag{3.43}$$

The claim follows by expanding the square and applying the marginal constraints. □

Lastly, we recall that the optimal transport distance metrizes narrow convergence, i.e.

Proposition 3.6 ([115], Theorem 5.11)**.** *On (subsets of) $\mathbb{R}^d$, $W_2(\rho_n,\rho) \to 0$ if and only if $\rho_n \rightharpoonup \rho$ and $\int |x|^2 \mathrm{d}\rho_n \to \int |x|^2 \mathrm{d}\rho$.*

Remark 3.13. *Note that the narrow convergence fulfills a number of criteria that other notions of convergence do not. For example, weak-$*$ convergence, in duality with $\mathcal{C}_0$ allows loss of mass at the boundary. Consider $\Omega = \mathbb{R}$ and a sequence of Dirac measures $\{\delta_{x_n}\}_n$ where $x_n \to +\infty$. Then $\int \varphi\delta_{x_n} \to 0$ for all $\varphi \in \mathcal{C}_0$ and the limit is not even a probability measure.*

3.6 Dynamical formulation

Given an optimal transport plan T, we know for every $x \in \operatorname{supp}(\rho)$ its destination $y \in \operatorname{supp}(\sigma)$. We can visualize this transport by introducing a time-like variable t:

Definition 3.7 (Displacement interpolation). *If the assumptions of Brenier's Theorem 3.2 are met, the curve*

$$t \mapsto \rho_t := \left((1-t)\mathrm{id} + tT_{\rho_0 \to \rho_1}\right)_\sharp \rho \tag{3.44}$$

is called the displacement interpolation *between ρ_0 and ρ_1, where $T_{\rho_0 \to \rho_1}$ is the optimal transport map connecting them.*

Remark 3.14. *The displacement formulation can more generally be defined for the transport plan π between ρ_0 and ρ_1 as $\rho_t := ((x,y) \mapsto (1-t)x + ty)_\sharp \pi$.*

As illustrated in Figure 3.2, the displacement interpolation between two measures resembles a transport process while the L^2 interpolation could be called a teleportation process at best.

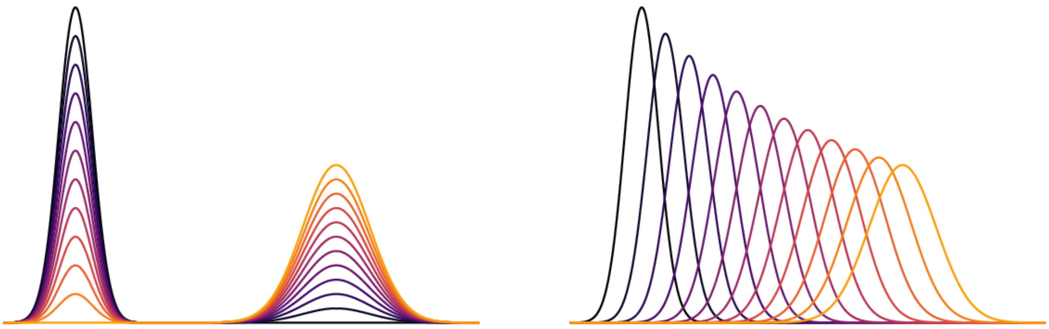

Figure 3.2: L^2 interpolation and displacement interpolation between two Gaussians.

Lagrangian picture

Note that if we track a single parcel of mass originating at x over time, it will move in a straight line $t \mapsto T_t(x)$ with constant speed $\partial_t T_t(x) = T(x) - x$ towards its destination. In the language of Lagrangian fluid mechanics, this vector field

$$v_{\mathrm{OT}}(t, T_t(x)) := T(x) - x \tag{3.45}$$

generates the *flow* of particles. Note that this is for now purely a formal statement, as we do not discuss if the vector field satisfies a Lipschitz condition to guarantee the well-posedness of the Cauchy problem defining the flow, which is given, for an arbitrary vector field $v : [0,1] \times \Omega \to \mathbb{R}^d$, by

$$\frac{\mathrm{d}}{\mathrm{d}t} F_t(x_0) = v(t, F_t(x_0)) \text{ with initial condition } F_0(x_0) = x_0 \ \forall x_0 \in \Omega. \tag{3.46}$$

Let us follow this line of thinking further and forget for now the optimal transport map and issues of regularity. Assume we are given a flow F_t that is moving our (smooth) initial density ρ_0 around Ω: $\rho_{F_t} := (F_t)_\sharp \rho_0$. We can obtain an evolution equation for ρ_{F_t} by differentiating this relation with respect to t, using the definition of the flow as well as

$$\frac{\mathrm{d}}{\mathrm{d}t} \det DF_t(x_0) = (\nabla \cdot v(t, F_t(x_0))) \det DF_t(x_0), \tag{3.47}$$

a special case of Liouville's formula. We find

$$\partial_t \rho_{F_t}(x) = \nabla \cdot (v(t,x)\rho_{F_t}(x)) \ \forall x \in \Omega \tag{3.48}$$

and recognize the *continuity equation.* Next, assume we want the flow to move $\rho_{F_0} = \rho_0$ to the configuration $\rho_{F_1} = \rho_1$.

Minimum energy flows

There are naturally many flows that do this. For the sake of uniqueness we choose a selection criterion. A natural one is to choose a flow of minimal action, given by the integral of the kinetic energy density $\frac{1}{2}|v(t,x)|^2 \rho_{F_t}(x)$. We obtain the following minimization problem:

$$\min_{v,\rho_{F_t}} \frac{1}{2} \int_0^1 \int_\Omega |v(t,x)|^2 \mathrm{d}\rho_{F_t}(x) \mathrm{d}t \tag{3.49}$$

subject to $\partial_t \rho_{F_t}(x) = \nabla \cdot (v(t,x)\rho_{F_t}(x))$ and $\rho_{F_0} = \rho_0, \rho_{F_1} = \rho_1$.

By the definition of the flow and its generating vector field, $F_0 = \mathrm{id}$, Fubini's theorem, and Jensen's inequality, we can quickly find a lower bound on this quantity:

$$\begin{aligned} \frac{1}{2} \int_0^1 \int_\Omega |v(t,x)|^2 \mathrm{d}\rho_{F_t}(x)\mathrm{d}t &= \frac{1}{2} \int_0^1 \int_\Omega \left| \frac{\mathrm{d}}{\mathrm{d}t} F_t(x_0) \right|^2 \mathrm{d}\rho_0(x_0)\mathrm{d}t & (3.50) \\ &\geq \frac{1}{2} \int_\Omega \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} F_t(x_0) \mathrm{d}t \right|^2 \mathrm{d}\rho_0 & (3.51) \\ &= \frac{1}{2} \int_\Omega |F_1(x_0) - x_0|^2 \mathrm{d}\rho_0 & (3.52) \end{aligned}$$

In fact, we know that the inequality in Equation (3.50) is an equality if and only if $\mathrm{d}F_t/\mathrm{d}t$ is constant in t. But we also know that this is the case for the choice $v(t, F_t(x_0)) = F_1(x_0) - x_0$.

The remarkable conclusion is that a solution of the minimization problem given by Equation (3.49) is given by the displacement interpolation ρ_t and its vector field

$$v_{\mathrm{OT}} : (t, T_t(x)) \mapsto T(x) - x, \tag{3.53}$$

Since the energy we minimize is convex, it is straightforward to show that this minimizer is unique.

In the Eulerian description, v_{OT} satisfies

$$\partial_t v_{\mathrm{OT}} + v_{\mathrm{OT}} \cdot \nabla v_{\mathrm{OT}} = 0, \tag{3.54}$$

the pressure-less Euler equation.

If v_{OT} is indeed the minimizer of Equation (3.49), then variations of the energy have to vanish at v_{OT}. Note that the variations, which we assume to be localized in time, have to be chosen such that the continuity equation is still fulfilled, that is if we let $v_\epsilon = v_{\mathrm{OT}} + \epsilon v'$, then necessarily $\nabla \cdot (v'\rho_t) = 0$. Hence, this quantity is a divergence-free vector field. Since

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} \int_\Omega \frac{1}{2}|v_{\mathrm{OT}} + \epsilon v'|^2 \mathrm{d}\rho_t = \int_\Omega v_{\mathrm{OT}} \cdot (v'\rho_t)\mathrm{d}x \overset{!}{=} 0, \tag{3.55}$$

the vector field v_{OT} is L^2-orthogonal to arbitrary divergence-free vector fields. Consequently, it is a gradient field. This ends our formal calculations and reproduces the result from Brenier's theorem, which followed then from the relation of the c-transform and Legendre transform for the quadratic cost.

A rigorous formulation and proof of the correspondence sketched here can be found in [115], Theorems 5.14 and 5.28. The equivalent definition of the OT distance through this dynamical problem is known as the *Bennamou-Brenier formula*:

$$W_2(\rho,\sigma)^2 = \min_{v_t,\nu_t}\left(\int_0^1 \|v_t\|^2_{L^2(\nu_t)}\mathrm{d}t : \partial_t\nu_t + \nabla \cdot (v_t\nu_t) = 0, \nu_0 = \rho, \nu_1 = \sigma\right). \tag{3.56}$$

3.7 The tangent space of $(\mathcal{P}(\Omega), W_2)$

The pair $(P(\Omega), W_2)$ forms a metric space, commonly referred to as the *Wasserstein-2 space*. The following result, Theorem 5.27 in [115], shows that the shortest path from ρ_0 to ρ_1 in this sense is given by the displacement interpolation:

Proposition 3.7. *The displacement interpolation between $\rho_0, \rho_1 \in \mathcal{P}(\Omega)$ is the constant speed geodesic connecting them, i.e.*

$$W_2(\rho_t,\rho_s)^2 = |t-s|^2 W_2(\rho_0,\rho_1)^2 \ \forall t,s \in [0,1]. \tag{3.57}$$

We have a notion of Lipschitz continuity and also the metric derivative of a curve $t \mapsto \rho(t) \in \mathcal{P}(\Omega)$, given by

$$|\dot\rho|(t) := \lim_{h\to 0}\frac{1}{h}W_2(\rho(t+h),\rho(t)). \tag{3.58}$$

It is possible to relate to every absolutely continuous (see [115], Section 5.3) curve $t \mapsto \rho_t$ through $(\mathcal{P}(\Omega), W_2)$ a vector field $v_t \in L^2(\Omega, \mu_t, \mathbb{R}^d)$ such that the continuity equation $\partial_t\rho_t + \nabla \cdot (\rho_t v_t) = 0$ is satisfied weakly. A proof of this can be found in [115], Theorem 5.14 for compact Ω and [8], Theorem 2.29 for $\Omega = \mathbb{R}^d$. We have already seen the formal considerations in the previous section, but this result in its generality assumes no regularity for the densities ρ_t. By the variational argument sketched in Equation (3.55), one knows that those v_t of minimal $L^2(\rho_t)$ norm are in fact gradients.

This space of vector fields is linear, and it is natural to call it the *tangent space* of $\mathcal{P}(\Omega)$. Indeed, we recall here Definition 2.31 from [8]:

Definition 3.8 (Tangent space of $(\mathcal{P}(\Omega), W_2)$)**.** *The tangent space of $(\mathcal{P}(\Omega), W_2)$ at $\rho \in \mathcal{P}(\Omega)$ is given by*

$$T_\rho\mathcal{P}(\Omega) := \left\{v \in L^2(\Omega,\rho,\mathbb{R}^d) : \int v \cdot w \,\mathrm{d}\rho = 0 \ \forall w \in L^2(\Omega,\rho,\mathbb{R}^d) : \nabla \cdot (w\rho) = 0\right\}. \tag{3.59}$$

At the same time, formally and geometrically, the tangent space of $\mathcal{P}(\Omega)$ at ρ_0, denoted $T_{\rho_0}\mathcal{P}(\Omega)$, is the set of tangent vectors to (smooth) curves

$$\rho : (-\epsilon, \epsilon) \to \mathcal{P}(\Omega), t \mapsto \rho_t \tag{3.60}$$

passing through ρ_0. If ρ_0 is a smooth, strictly positive density on Ω, then this space contains functions $\varphi : \Omega \to \mathbb{R}$ with $\int_\Omega \varphi = 0$, as $\rho_0 + \epsilon\varphi$ remains a probability density for suitably small ϵ. This notion of tangent space is introduced in [102] and sketched in Figure 3.3. In the case where transport is induced by maps, the two notions of tangent spaces coincide and one can move from one to the other by taking $\nabla\psi$ and considering the curve $t \mapsto (\mathrm{id} - t\nabla\psi)_\sharp\rho_0$ since

$$\begin{aligned}
\partial_t\rho_t\big|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} (\mathrm{id} - t\nabla\psi)_\sharp\rho_0(x) && (3.61)\\
&= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \big(\rho_0(x - t\nabla\psi(x)) \det(\mathrm{Id} - tD^2\psi(x))\big) && (3.62)\\
&= -\nabla \cdot (\rho_0\nabla\psi). && (3.63)
\end{aligned}$$

We refer to [8], Section 6.2, for a detailed description of the two concepts and how they relate in more general cases.

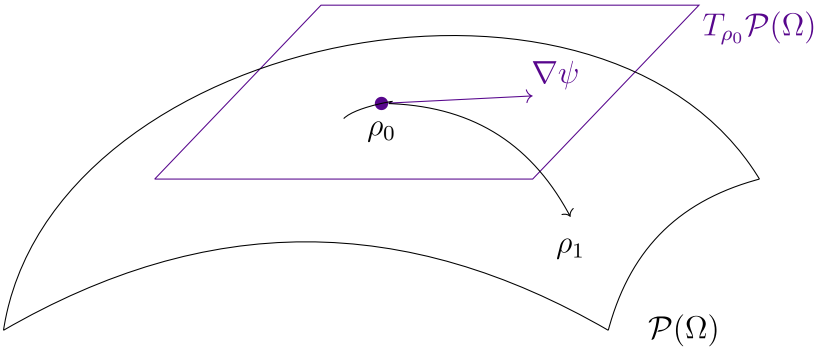

Figure 3.3: The geometric tangent space of $\mathcal{P}_{\mathrm{ac}}(\Omega)$.

We will continue on a formal level with using these ideals to consider a Riemannian structure on $(\mathcal{P}(\Omega), W_2)$, following the work of Otto in [102]. On $T_{\rho_0}\mathcal{P}(\Omega)$, introduce the following inner product between tangent vectors to two curves $t \mapsto \rho_t^{(1)}, t \mapsto \rho_t^{(2)}$ passing through ρ_0:

$$\langle \partial_t\rho^{(1)}, \partial_t\rho^{(2)} \rangle_{\rho_0} := \int \nabla\psi^{(1)} \cdot \nabla\psi^{(2)} \, \mathrm{d}\rho_0, \tag{3.64}$$

where $\partial_t\big|_{t=0}\rho^{(i)} + \nabla \cdot (\rho_0\nabla\psi^{(i)}) = 0$ for $i = 1, 2$. Note that by comparison with Equation (3.56), this implies that

$$W_2(\rho, \sigma)^2 = \min_{\nu_t} \left(\int_0^1 \|\partial_t\nu_t\|_{\nu_t}^2 \mathrm{d}t : \nu_0 = \rho, \nu_1 = \sigma \right). \tag{3.65}$$

Following [102], this choice of inner product on $T_{\rho_0}\mathcal{P}(\Omega)$ implies that the push-forward operation of the measure ρ_0 is an *isometric submersion* from $(\mathrm{Diff}(\Omega), \langle\cdot,\cdot\rangle_F)$ into $(\mathcal{P}(\Omega), \langle\cdot,\cdot\rangle_{\rho_0})$. The space $\mathrm{Diff}(\Omega)$ consists of diffeomorphisms $F : \Omega \to \Omega$. Again formally, the tangent space of $\mathrm{Diff}(\Omega)$ can be identified with smooth vector fields on Ω that are tangent to its boundary.

Remark 3.15. *An idea popularized by Arnold [10] is to interpret the equations of fluid mechanics as curves of shortest length on* $\mathrm{Diff}(\Omega)$. *When one considers incompressible fluid mechanics, these diffeomorphisms have to be measure-preserving, i.e. $S_\sharp\rho = \rho$, which implies that their corresponding vector fields are divergence-free in the sense that $\nabla \cdot ((v \circ S^{-1})\rho) = 0$. This makes them orthogonal to those from optimal transport in the sense of the* polar decomposition theorem.

More on this connection is presented in Appendix B.

3.8 Monge embeddings

Let us fix a reference density $\bar\rho \in \mathcal{P}(\Omega)$. If we follow the geometric interpretation from the previous section further, the map that sends ρ to $T_{\bar\rho\to\rho}$ is the right inverse of the exponential map in Riemannian geometry: it takes an element of the non-linear space $\mathcal{P}(\Omega)$ and assigns to it an element of the linear space $T_{\bar\rho}\mathcal{P}(\Omega) \subset L^2(\bar\rho, \mathbb{R}^d)$.

Remark 3.16. *The exponential map itself is given by the push-forward operation that assigns to a vector field $v \in L^2(\Omega, \rho, \mathbb{R}^d)$ the element $(\mathrm{id} - v)_\sharp\bar\rho \in \mathcal{P}(\Omega)$. From the inequality*

$$W_2((\mathrm{id} - v^{(1)})_\sharp\bar\rho, (\mathrm{id} - v^{(2)})_\sharp\bar\rho)^2 \leq \int |v^{(1)} - v^{(2)}|^2 \mathrm{d}\bar\rho, \tag{3.66}$$

(which follows from taking the trial plan $((\mathrm{id} - v^{(2)}, \mathrm{id} - v^{(1)})^{-1})_\sharp\bar\rho$), we can conclude that the exponential map is non-expansive. This hints at the fact that $(\mathcal{P}(\Omega), W_2)$ has negative curvature [66].

Definition 3.9 (Monge embedding [94, 141])**.** *Given a reference density $\bar\rho \in \mathcal{P}(\Omega)$, absolutely continuous with respect to the Lebesgue measure, we call*

$$\mathrm{ME}_{\bar\rho} : \mathcal{P}(\Omega) \to L^2(\Omega, \rho, \mathbb{R}^d), \rho \mapsto T_{\bar\rho\to\rho} \tag{3.67}$$

the Monge embedding *with respect to $\bar\rho$.*

Remark 3.17. *We will omit the reference density $\bar\rho$ when there is no risk of ambiguity and write T_ρ for $T_{\bar\rho\to\rho}$. In these cases, we denote φ_ρ the convex function such that $\nabla\varphi_\rho = T_\rho$ and ψ_ρ the potential such that $T_\rho = \mathrm{id} - \nabla\psi_\rho$. We enforce uniqueness of the potential by letting $\int \varphi_\rho \mathrm{d}\rho = 0$.*

We recall some properties of the Monge embedding.

Proposition 3.8. *The Monge embedding is continuous.*

This is a result of the *stability of optimality.* For a sequence of densities ρ_k narrowly converging to ρ, if corresponding optimal transport maps T_k between ρ_k and a given $\bar\rho$ exist, and if there exists a unique optimal transport map from ρ to $\bar\rho$, denoted T, then the T_k converge to T in measure, relative to $\bar\rho$ ([139], Corollary 5.23).

Proposition 3.9. *For any measures ρ_1, ρ_2,*

$$W_2(\rho_1, \rho_2) \leq \|\mathrm{ME}_{\bar\rho}(\rho_1) - \mathrm{ME}_{\bar\rho}(\rho_2)\|_{L^2(\bar\rho)} = \|T_{\rho_1} - T_{\rho_2}\|_{L^2(\bar\rho)}. \tag{3.68}$$

Proof. The transport plan $(T_{\rho_1}, T_{\rho_2})_\sharp \bar\rho$ is an admissible competitor in $\Pi(\rho_1, \rho_2)$ and

$$W_2^2(\rho_1, \rho_2) \leq \int |x-y|^2 \mathrm{d}((T_{\rho_1}, T_{\rho_2})_\sharp \bar\rho)(x,y) = \int |T_{\rho_1}(x) - T_{\rho_2}(x)|^2 \mathrm{d}\bar\rho(x). \tag{3.69}$$

□

Hölder continuity

Next, we give a result improving Proposition 3.8, taken from [66] and attributed to Ambrosio.

Proposition 3.10. *Let $[0,1] \ni t \mapsto \rho_t$ be a Lipschitz curve with constant L through $(\mathcal{P}(\Omega), W_2)$. Assume that $\bar\rho$ and ρ_0 satisfy the assumptions of Theorem 6.1 (Sufficient conditions for strict convexity of the transport map). Denote by T_t the optimal transport map from $\bar\rho$ to ρ_t for $0 \leq t \leq 1$. Then,*

$$\limsup_{t\to 0} \frac{\|T_t - T_0\|_{L^2(\bar\rho)}}{\sqrt{t}} < +\infty. \tag{3.70}$$

We will repeat the proof here because it is quite instructive.

Lemma 3.1. *Let ρ and σ as in Theorem 3.5 (Caffarelli's regularity result) and let denote by λ the modulus of convexity of $\varphi_{\rho\to\sigma}$, i.e. the largest $\lambda' > 0$ such that $x \mapsto \varphi_{\rho\to\sigma}(x) - \lambda' \frac{|x|^2}{2}$ is convex on the support of ρ. Let $T_{\sigma\to\rho} = (\nabla\varphi_{\rho\to\sigma})^{-1}$. Then, for any map $S_{\sigma\to\rho}$, it holds that*

$$\|S_{\sigma\to\rho} - T_{\sigma\to\rho}\|^2_{L^2(\sigma)} \leq \frac{2}{\lambda}\left(\|S_{\sigma\to\rho} - \mathrm{id}\|^2_{L^2(\sigma)} - \|T_{\sigma\to\rho} - \mathrm{id}\|^2_{L^2(\sigma)}\right). \tag{3.71}$$

Proof. We omit the subscripts from T and S for brevity. Note that $\int \varphi(x)\mathrm{d}\rho(x) = \int \varphi(S(x))\mathrm{d}\sigma(x) = \int \varphi(T(x))\mathrm{d}\sigma(x)$ by definition. From the strict convexity of φ, we obtain

$$\int (\nabla\varphi \circ T)\cdot(S-T)\mathrm{d}\sigma + \frac{1}{2}\lambda\|S-T\|^2_{L^2(\sigma)} \geq 0. \tag{3.72}$$

On $\mathrm{supp}(\sigma)$, $\varphi \circ T(x) = x$ by definition, hence

$$\int (\nabla\varphi \circ T(x))\cdot(S(x)-T(x))\mathrm{d}\sigma(x) = \int x\cdot(S(x)-T(x))\mathrm{d}\sigma(x) \tag{3.73}$$

$$= \frac{1}{2}\|S-\mathrm{id}\|^2_{L^2(\sigma)} - \frac{1}{2}\|T-\mathrm{id}\|^2_{L^2(\sigma)} \tag{3.74}$$

and the claim follows. □

Proof of Proposition 3.10. Denote by S_t the optimal transport map from ρ_t to ρ_0. Both T_0 and $S_t \circ T_t$ map $\bar\rho$ into ρ_0. From Lemma 3.1, we obtain

$$\|S_t \circ T_t - T_0\|^2_{L^2(\bar\rho)} \leq C\left(\|S_t \circ T_t - \mathrm{id}\|^2_{L^2(\bar\rho)} - \|T_0 - \mathrm{id}\|^2_{L^2(\bar\rho)}\right) \tag{3.75}$$

with $C = \frac{2}{\lambda_0}$, depending on the convexity of T_0 but not on t. Next,

$$\begin{aligned}
\|S_t \circ T_t - \mathrm{id}\|_{L^2(\bar\rho)} &\leq \|S_t \circ T_t - T_t\|_{L^2(\bar\rho)} + \|T_t - \mathrm{id}\|_{L^2(\bar\rho)} && (3.76)\\
&= \|S_t - \mathrm{id}\|_{L^2(\rho_t)} + W_2(\rho_t, \bar\rho) && (3.77)\\
&= W_2(\rho_t, \rho_0) + W_2(\rho_t, \bar\rho) && (3.78)\\
&\leq 2W_2(\rho_t, \rho_0) + W_2(\rho_0, \bar\rho) && (3.79)\\
&\leq 2Lt + W_2(\rho_0, \bar\rho) && (3.80)
\end{aligned}$$

At the same time,

$$\begin{aligned}
\|S_t \circ T_t - T_0\|_{L^2(\bar\rho)} &\geq \|T_t - T_0\|_{L^2(\bar\rho)} - \|S_t \circ T_t - T_t\|_{L^2(\bar\rho)} && (3.81)\\
&= \|T_t - T_0\|_{L^2(\bar\rho)} - \|S_t - \mathrm{id}\|_{L^2(\rho_t)} && (3.82)\\
&\geq \|T_t - T_0\|_{L^2(\bar\rho)} - Lt. && (3.83)
\end{aligned}$$

With this, Equation (3.75) becomes

$$\begin{aligned}
(\|T_t - T_0\|_{L^2(\bar\rho)} - Lt)^2 &\leq C((2Lt + W_2(\rho_0, \bar\rho))^2 - W_2(\rho_0, \bar\rho)^2) && (3.84)\\
&= C(4L^2t^2 + 2LtW_2(\rho_0, \bar\rho)) && (3.85)
\end{aligned}$$

The claim follows. □

There exist a number of explicit examples that show that $\frac{1}{2}$-Hölder regularity is the best one can expect in general ([66], Section 4).

In the case where ρ, σ do not enjoy the regularity conditions assumed in the previous proposition, the following results hold.

Proposition 3.11 ([19], Proposition 3.4, [94], Theorem 3.1)**.** *For $\rho_1, \rho_2 \in \mathcal{P}(\Omega)$ on a compact, convex $\Omega \subset \mathbb{R}^d$ with unit volume and $\bar\rho \equiv 1$,*

$$\|T_{\rho_1} - T_{\rho_2}\|_{L^2(\bar\rho)} \leq CW_1(\rho_1, \rho_2)^{2/15} \tag{3.86}$$

and

$$\|T_{\rho_1} - T_{\rho_2}\|_{L^2(\bar\rho)} \leq CW_1(\rho_1, \rho_2)^{1/(2^{d-1}(d+2))}, \tag{3.87}$$

where the constants depend only on Ω.

Compatible measures and maps

As was already hinted at in the proof of Proposition 3.10, the difference between $W_2(\rho_1, \rho_2)$ and $\|T_{\bar\rho\to\rho_1} - T_{\bar\rho\to\rho_2}\|_{L^2(\bar\rho)}$ is introduced by linking them via $\bar\rho$ and the triangle inequality. This is most obvious if $T_{\bar\rho\to\rho_2}$ is invertible:

$$\begin{aligned}
\|T_{\bar\rho\to\rho_1} - T_{\bar\rho\to\rho_2}\|_{L^2(\bar\rho)} - W_2(\rho_1, \rho_2) &= \|T_{\bar\rho\to\rho_1} \circ T_{\rho_2\to\bar\rho} - \mathrm{id}\|_{L^2(\rho_2)} - \|T_{\rho_2\to\rho_1} - \mathrm{id}\|_{L^2(\rho_2)} && (3.88)\\
&\leq \|T_{\bar\rho\to\rho_1} \circ T_{\rho_2\to\bar\rho} - T_{\rho_2\to\rho_1}\|_{L^2(\rho_2)}, && (3.89)
\end{aligned}$$

The following proposition is a direct consequence.

Proposition 3.12. *For $\rho_1, \rho_2, \bar\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, the Monge embedding is an isometry if*

$$T_{\bar\rho\to\rho_1} = T_{\rho_2\to\rho_1} \circ T_{\bar\rho\to\rho_2} \tag{3.90}$$

Definition 3.10 (Compatibility between Monge embedding and push-forward [91])**.** *We say that the Monge embedding with reference density $\bar{\rho}$, a measure $\sigma \in \mathcal{P}(\Omega)$, and a map $F \in L^2(\Omega, \mathbb{R}^d, \bar{\rho})$ are* compatible *if*

$$T_{\bar{\rho}\to F_\sharp\sigma} = F \circ T_{\bar{\rho}\to\sigma} \quad \textit{i.e.} \quad \mathrm{ME}_{\bar{\rho}}(F_\sharp\sigma) = F \circ \mathrm{ME}_{\bar{\rho}}(\sigma). \tag{3.91}$$

Moreover, we say that a set of measures $\{\rho_i\}_{i=1}^m \subset \mathcal{P}_{\mathrm{ac}}(\Omega)$ is compatible if

$$T_{\rho_j\to\rho_k} \circ T_{\rho_i\to\rho_j} = T_{\rho_i\to\rho_k} \tag{3.92}$$

for all $1 \leq i, j, k \leq m$.

Remark 3.18. *If we denote by $\sharp_{\bar{\rho}}$ the operation that takes a map $F \in L^2(\Omega, \mathbb{R}^d, \bar{\rho})$ to $F_\sharp\bar{\rho} \in \mathcal{P}(\Omega)$, for any $\sigma \in \mathcal{P}(\Omega)$,*

$$\sharp_{\bar{\rho}} \circ \mathrm{ME}_{\bar{\rho}}(\sigma) = (T_{\bar{\rho}\to\sigma})_\sharp\bar{\rho} = \sigma. \tag{3.93}$$

In contrast,

$$\mathrm{ME}_{\bar{\rho}} \circ \sharp_{\bar{\rho}}(F) = T_{\bar{\rho}\to F_\sharp\bar{\rho}} \tag{3.94}$$

is only equal to F if F itself is an optimal transport map, i.e. a gradient $\in T_{\bar{\rho}}\mathcal{P}(\Omega)$. This is in agreement with the Riemannian picture of optimal transport, where $\sharp_{\bar{\rho}}$ plays the role of the exponential map. The condition $\mathrm{ME}_{\bar{\rho}} \circ \sharp_{\bar{\rho}}(F) = F$ is a special case of Equation (3.91) *for $\sigma = \bar{\rho}$.*

The compatibility condition is in particular fulfilled by maps T that have the form of shifts and scalings:

Proposition 3.13. *If $\rho_1, \rho_2 \in \mathcal{P}(\Omega)$ are related by shifts and scalings, i.e. there exists $T_{\rho_2\to\rho_1} : x \mapsto sx + a$ with $a \in \mathbb{R}^d$ and $s \in \mathbb{R}_{>0}$, then Equation* (3.90) *holds.*

Proof. It is enough to show that the map $T_{\rho_2\to\rho_1} \circ T_{\bar{\rho}\to\rho_2}$ is the gradient of a convex function, as it satisfies the push-forward condition. This is clear from the assumptions, in particular

$$T_{\rho_2\to\rho_1} \circ T_{\bar{\rho}\to\rho_2}(x) = \nabla(s\varphi_{\bar{\rho}\to\rho_2}(x) + a \cdot x). \tag{3.95}$$

□

In [3], it is shown that this is in general a necessary condition in the following sense:

Proposition 3.14 ([3], Theorem 4.4)**.** *Let $d > 1$ and consider $\mathcal{F} := \{\nabla\varphi : \varphi \in \mathcal{C}^2(\mathbb{R}^d)$ strictly convex$\}$. Furthermore, if for any $\sigma \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ with compact support, it holds that*

$$T_{\bar{\rho}\to F_\sharp\sigma} = F \circ T_{\bar{\rho}\to\sigma} \tag{3.96}$$

for all $F = \nabla\chi$ in a subset of $\mathcal{F}$, then all of these F have the form of shifts and scalings.

The proof of this proposition follows directly from the following result: Without any further restriction on the form of $\nabla\varphi_{\bar{\rho}\to\sigma}$, the function $\nabla\chi \circ \nabla\varphi_{\bar{\rho}\to\sigma}$ can only be written as the gradient of another strictly convex function $\varphi_{\bar{\rho}\to\nabla\chi_\sharp\sigma}$ if $D^2\chi \equiv s\mathrm{Id}$ for some $s \in \mathbb{R}_{>0}$ ([3], Proposition 4.12).

The Monge embedding is very intriguing, as it allows us to move from the nonlinear $(\mathcal{P}(\Omega), W_2)$ to the linear $L^2(\Omega, \bar{\rho}, \mathbb{R}^d)$ while retaining information about translations and uniform dilations. Note that this was one of the main reasons we introduced the optimal transport framework in this work in the first place, c.f. Figure 3.2. Furthermore, the reference density $\bar{\rho}$ is free to choose, and we can always choose one that admits a density.

Hilbertian distances

However, it should be noted that no matter how well-chosen it is, the Monge embedding can not encode all information of W_2, as the Wasserstein distance is not *Hilbertian* for $d > 1$.

Definition 3.11 (Hilbertian distance). *We say that a distance* dist *on a set* Σ *is Hilbertian if there exists a map* $\mathfrak{F} : \Sigma \to V$*, where* V *is a Hilbert space and* $\mathrm{dist}(s, s') = \|\mathfrak{F}(s) - \mathfrak{F}(s')\|_V \ \forall s, s' \in \Sigma$.

Remark 3.19. *We call the map* $\mathfrak{F}$ *a* feature map*. Hilbertian distances are of great interest in data science applications as they form the basis for Kernel methods, where one can use the properties of the feature space* V *without ever having to evaluate the map* $\mathfrak{F}$ *explicitly.*

Proposition 3.15 ([106], Proposition 8.1). dist *is Hilbertian if and only if* dist^2 *is negative definite.*

We refer to the references for a proof and state the following corollary:

Remark 3.20. *If* $\mathbb{D}$ *is a* $n_s \times n_s$ *matrix with entries* $\{W_2(\rho_i, \rho_j)^2\}_{1 \le i,j \le n_s}$ *for a set of measures* $\{\rho_i\}_i$*, then the centered distance matrix* ($\mathbb{A}\mathbb{D}\mathbb{A}$ *where* $\mathbb{A}_{ij} = \delta_{ij} - \frac{1}{n_s}$ *is not negative definite. In contrast, if* $\mathbb{D}_V$ *is the matrix of distances* $\|u_i - u_j\|_V^2$ *for* $\{u_i\}_i \subset V$*, a finite-dimensional Hilbert space, then* $\mathbb{A}\mathbb{D}_V\mathbb{A}$ *corresponds to the Gram matrix with entries* $\langle u_i, u_j \rangle_V : 1 \le i, j \le n_s$*, from which one can re-construct the positions* u_i *up to a global translation and rotation [52].*

Almost compatible maps and measures

At this point, it is worth to revisit two assumptions we have taken so far: If Ω is assumed to be bounded and $\bar{\rho}$ is assumed to be strictly positive in the entire domain (to apply the regularity results of Theorem 3.5), then $T_{\bar{\rho}\to\rho}$ is not going to correspond to a pure shifting and scaling operation. However, we can imagine a situation where the transport is ϵ-close to a shift-and scaling operation in the sense that $\|T_{\bar{\rho}\to\rho} - S\|_{L^2(\bar{\rho})} \le \epsilon$ where S is of the form $S(x) = sx + a$ as before.

For such maps which are ϵ-close to shifting and scaling maps, the following result is shown in [91], Theorem 4.1:

Proposition 3.16. *Let $\bar{\rho}, \sigma \in \mathcal{P}(\Omega)$, $\bar{\rho}$ absolutely continuous, $R, \epsilon > 0$, and*

$$\begin{aligned}\mathcal{G}_{\sigma,R,\epsilon} := \{G \in L^2(\bar{\rho}, \mathbb{R}^d) : \exists F \in L^2(\bar{\rho}, \mathbb{R}^d), a \in \mathbb{R}^d, s \in \mathbb{R}_{>0} :& \\ F(x) = sx + a, \|F\|_{L^2(\sigma)} \leq R, \text{ and } \|G - F\|_{L^2(\sigma)} \leq \epsilon\}.& \quad (3.97)\end{aligned}$$

Then, for $\bar{\rho} \equiv |\Omega|^{-1}$ with $\Omega \subset \mathbb{R}^d$ convex and compact, and $G_1, G_2 \in \mathcal{G}_{\sigma,R,\epsilon}$:

$$0 \leq \|\mathrm{ME}_{\bar{\rho}}((G_1)_\sharp\sigma) - \mathrm{ME}_{\bar{\rho}}((G_2)_\sharp\sigma)\|_{L^2(\bar{\rho})} - W_2((G_1)_\sharp\sigma, (G_2)_\sharp\sigma) \leq C\epsilon^{2/15} + 2\epsilon. \quad (3.98)$$

Furthermore, if $\bar{\rho}$ and σ satisfy Caffarelli's regularity assumptions of Theorem 3.5, then

$$0 \leq \|\mathrm{ME}_{\bar{\rho}}((G_1)_\sharp\sigma) - \mathrm{ME}_{\bar{\rho}}((G_2)_\sharp\sigma)\|_{L^2(\bar{\rho})} - W_2((G_1)_\sharp\sigma, (G_2)_\sharp\sigma) \leq C'\epsilon^{1/2} + C\epsilon. \quad (3.99)$$

All constants depend on $\bar{\rho}, \sigma$, and R.

Proof. We will show only the second case, which can be proved using Proposition 3.10. Let F_1 be a compatible map ϵ-close to G_1. Let $(F_i)_\sharp\sigma =: \rho_i^\epsilon$ and $(G_i)_\sharp\sigma =: \rho_i$ for $i = 1, 2$.

First, note that since the transport via σ is an admissible competitor plan,

$$\epsilon > \|G_i - F_i\|_{L^2(\sigma)} \geq W_2(\rho_i, \rho_i^\epsilon) \quad \forall i \in \{1, 2\}. \quad (3.100)$$

Second, by the triangle inequality,

$$\begin{aligned}&\|T_{\bar{\rho}\to\rho_1} - T_{\bar{\rho}\to\rho_2}\|_{L^2(\bar{\rho})} \\ &\quad \leq \|T_{\bar{\rho}\to\rho_1} - T_{\bar{\rho}\to\rho_1^\epsilon}\|_{L^2(\bar{\rho})} + \|T_{\bar{\rho}\to\rho_2} - T_{\bar{\rho}\to\rho_2^\epsilon}\|_{L^2(\bar{\rho})} + \|T_{\bar{\rho}\to\rho_1^\epsilon} - T_{\bar{\rho}\to\rho_2^\epsilon}\|_{L^2(\bar{\rho})} \quad (3.101)\end{aligned}$$

By the compatibility of $(\mathrm{ME}_{\bar{\rho}}, \sigma, F_i)$, the last term is bounded by

$$\|T_{\bar{\rho}\to\rho_1^\epsilon} - T_{\bar{\rho}\to\rho_2^\epsilon}\|_{L^2(\bar{\rho})} \leq 2\epsilon + W_2(\rho_1, \rho_2), \quad (3.102)$$

using Equation (3.100). The last step is to bound $\|T_{\bar{\rho}\to\rho_i} - T_{\bar{\rho}\to\rho_i^\epsilon}\|_{L^2(\bar{\rho})}$ using Equation (3.75) where ρ_i plays the role of ρ_0 and ρ_i^ϵ that of ρ_t. For $i = 1$, we arrive at

$$\|T_{\bar{\rho}\to\rho_1} - T_{\bar{\rho}\to\rho_1^\epsilon}\|_{L^2(\bar{\rho})} - \underbrace{W_2(\rho_1, \rho_1^\epsilon)}_{\leq\epsilon} \leq \sqrt{\frac{4}{\lambda_{\rho_1^\epsilon}}} \left(W_2(\rho_1, \rho_1^\epsilon) + \sqrt{W_2(\rho_1^\epsilon, \rho_1) W_2(\rho_1^\epsilon, \bar{\rho})}\right), \quad (3.103)$$

where $\lambda_{\rho_1^\epsilon}$ is the modulus of convexity of the transport potential from $\bar{\rho}$ to ρ_1^ϵ. To conclude, we need to show that $\lambda_{\rho_1^\epsilon}$ and $W_2(\rho_1^\epsilon, \bar{\rho})$ are bounded by constants independent of F_1 and G_1. For this, we use the assumption that $\|F\|_{L^2(\sigma)} \leq R$:

$$W_2(\rho_1^\epsilon, \bar{\rho}) \leq W_2(\sigma, \bar{\rho}) + W_2(\sigma, \rho_1^\epsilon) = W_2(\sigma, \bar{\rho}) + \|F - \mathrm{id}\|_\sigma \leq W_2(\sigma, \bar{\rho}) + R + \|\mathrm{id}\|_\sigma. \quad (3.104)$$

We can bound $\lambda_{\rho_1^\epsilon}$ using λ_σ, the modulus of convexity of the transport potential from $\bar{\rho}$ to σ, as the two are connected by the shift and scaling operation F_1. The shifting operation does not change the modulus of convexity, while the scaling with s implies $1/\lambda_{\rho_1^\epsilon} = s/\lambda_\sigma < R/(\lambda_\sigma\|\mathrm{id}^2\|_\sigma)$ ([91], Corollary 6.6). □

3.9 Closed forms

There are some cases where the optimal transport problem simplifies to the point where explicit formulas are available for its computation.

One dimension

We have already seen in Equation (3.35) that if $\Omega \subset \mathbb{R}$, it holds that

$$T_{\rho\to\sigma} = \mathrm{cdf}(\sigma)^{[-1]} \circ \mathrm{cdf}(\rho).$$

This is a special case of a Monge embedding. Since the composition of two non-decreasing functions is always non-decreasing, we can conclude that all measures are compatible with each other in one spatial dimension. We can thus choose $\mathbb{1}_{[0,1]}$ as a reference and compute the optimal transport distance through

$$W_2(\rho,\sigma) = \|\mathrm{cdf}(\rho)^{[-1]} - \mathrm{cdf}(\sigma)^{[-1]}\|_{L^2([0,1])}. \tag{3.105}$$

Gaussian measures

Let ρ and σ are Gaussian measures on $\mathbb{R}^d$, with respective means m_ρ and m_σ as well as covariance matrices Var_ρ and Var_σ. They can be connecting by a shifting and (not necessarily uniform) scaling operation S. If one can show that this operation can be written as the gradient of a convex (in fact: quadratic) map, then $S = T$ is necessarily optimal.

By substituting a quadratic ansatz into the push-forward condition, one obtains (see, for example, [103], Section 1.6.3)

$$T_{\rho\to\sigma}(x) = m_\sigma + \mathrm{Var}_\rho^{-1/2}\left(\mathrm{Var}_\rho^{1/2}\mathrm{Var}_\sigma\mathrm{Var}_\rho^{1/2}\right)^{1/2}\mathrm{Var}_\rho^{-1/2}(x-m_\rho). \tag{3.106}$$

This also gives an explicit formula for the total transport cost, which serves as a lower bound for general measures.

Proposition 3.17 ([103], Proposition 1.6.5)**.** *Let $\rho,\sigma \in \mathcal{P}(\Omega)$ with respective means m_ρ, m_σ and covariance matrices $\mathrm{Var}_\rho, \mathrm{Var}_\sigma$. Then,*

$$W_2(\rho,\sigma)^2 \geq |m_\rho - m_\sigma|^2 + \mathrm{tr}(\mathrm{Var}_\rho + \mathrm{Var}_\sigma - 2(\mathrm{Var}_\rho^{1/2}\mathrm{Var}_\sigma\mathrm{Var}_\rho^{1/2})^{1/2}). \tag{3.107}$$

Furthermore, equality holds if ρ and σ are Gaussian.

Remark 3.21. *The trace expression in Equation* (3.107) *is known as the Bures metric on the space of symmetric positive semi-definite matrices.*

Remark 3.22. *The explicit form of T in the Gaussian case allows for a condition under which a set of Gaussian measures is compatible. Let $\bar\rho = \mathcal{N}(0,\mathrm{Id})$. Assume that $m_\rho = m_\sigma = 0$ without loss of generality (we know that optimal transport factors translations and that translated copies of a measure are compatible). Then,*

$$T_{\bar\rho\to\sigma}\circ T_{\bar\rho\to\rho} = T_{\rho\to\sigma} \Leftrightarrow \mathrm{Var}_\sigma^{1/2}\mathrm{Var}_\rho^{-1/2} = \mathrm{Var}_\rho^{-1/2}\left(\mathrm{Var}_\rho^{1/2}\mathrm{Var}_\sigma\mathrm{Var}_\rho^{1/2}\right)^{1/2}\mathrm{Var}_\rho^{-1/2}. \tag{3.108}$$

This condition is fulfilled if and only if Var_σ *and* Var_ρ *commute.*

Applicability of the Monge embedding approximation

The Monge embeddings provide a linearization of the Wasserstein space $(\mathcal{P}(\Omega), W_2)$ that retains some of the information encoded in the optimal transport distance. The cases where the embedding is an isometry are limited to $d = 1$ and other cases that are in essence one-dimensional, such as pure shifts and uniform scalings or cases where the dimensions are separable. The case of Gaussian measures with simultaneously diagonalizable covariance matrices can be seen as an example of the latter.

That said, the linear nature of the tangent space makes it possible to apply established linear separation methods on the Monge embeddings, which has been done for example in [91, 94, 141]. In these applications, the linear approximation was able to provide very good results, comparable to those obtained using the fully non-linear true optimal transport distance.

We will employ a similar strategy in Chapter 6, where we use a combination of a linear reduced basis approximation in a reference frame together with a registration map Φ_μ in order to represent solutions $u(\mu)$ of a PPDE problem. The mapping Φ_μ is built from Monge embeddings, which allows us to use linear dimension-reduction methods. The fact that the Monge embeddings come with a loss of information is not critical since we expect this loss of information to be representable in the reduced basis of the reference frame.

Chapter 4

Computing optimal transport

It is straightforward to think of a discrete counterpart to the optimal transport problem given in Definition 3.3. If we represent the distributions ρ and σ by a sum of weighted Dirac measures $\rho \approx \sum_{i=1}^{M} \hat{\rho}_i \delta_{x_i}$ and $\sigma \approx \sum_{j=1}^{M} \hat{\sigma}_j \delta_{y_j}$, then Equation (3.3) becomes

$$\min_{\hat{\pi} \in \mathbb{R}_{\geq 0}^{M \times M}} \sum_{i,j=1}^{M} \hat{c}_{ij} \hat{\pi}_{ij} : \sum_{j=1}^{M} \hat{\pi}_{ij} = \hat{\rho}_i \; \forall j = 1, \ldots, M \text{ and } \sum_{j=1}^{M} \hat{\pi}_{ij} = \hat{\sigma}_j \; \forall i = 1, \ldots, M, \tag{4.1}$$

where $\hat{c}_{ij} := c(x_i, y_j)$.

Remark 4.1. *We assume here for simplicity that ρ and σ are represented by the same number of Dirac measures. It is straightforward to extend all following considerations to the case where $\sigma \approx \sum_{j=1}^{M'} \hat{\sigma}_j \delta_{y_j}$ with $M' \neq M$.*

Equation (4.1) is a linear programming problem: a linear function is to be minimized with linear equality and inequality constraints. The challenge lies in the size of the problem: the transport plan is of size $M \times M$, which is much too large for practical values of M.

The dual problem Equation (3.5) is a more favorable starting point. In the discrete setting we just introduced it reads

$$\max_{\hat{\psi}_\rho, \hat{\psi}_\sigma \in \mathbb{R}^M} \left(\sum_{i=1}^{M} \hat{\psi}_{\rho,i} \hat{\rho}_i + \sum_{j=1}^{M} \hat{\psi}_{\sigma,j} \hat{\sigma}_j \right) : \hat{\psi}_{\rho,i} + \hat{\psi}_{\sigma,j} \leq \hat{c}_{ij} \; \forall 1 \leq i, j \leq M. \tag{4.2}$$

At first glance, it looks like the c-transform from Definition 3.4 provides us with a strategy to move up the dual problem in an iterative manner: start with an arbitrary initial potential ψ_ρ (e.g. identically zero), apply the c-transform to obtain ψ_ρ^c and so on.

However, as already noted, this strategy will stall quickly, since $\psi_\rho^{ccc} = \psi_\rho^c$, as already pointed out in Section 3.2. The way out is to relax the optimality condition to only enforce $\psi_\rho(x) + \psi_\sigma(y) \geq c(x, y) + \varepsilon$ for some $\varepsilon > 0$. The resulting method, called an *auction algorithm* provides potentials that are εN-close to optimality in (at its most naive implementation) $\varepsilon^{-1} N^3 \max_{\rho \otimes \sigma} c$ iterations [106], Section 3.7). We refer to [22] for a more detailed introduction to auction algorithms.

4.1 Entropic optimal transport

As it turns out, one can use a similar method of relaxation in order to obtain an iterative method that ascends the dual problem in, using the title of one of the most influential papers in this field, light speed [44].

Entropic optimal transport introduces a regularization term to the transport problem that relaxes the marginal constraint. This modification turns the primal problem strictly convex and the dual problem strictly concave. We describe it here first in its continuous form.

Primal and dual formulation

Definition 4.1 (Entropic optimal transport)**.** *Let $\rho, \sigma, c, \Pi(\rho, \sigma)$ as in Definition 3.3 and $\varepsilon > 0$. The OT problem with entropic regularization reads*

$$W^2_{2,\varepsilon}(\rho,\sigma)^2 := \min_{\pi^\varepsilon \in \Pi(\rho,\sigma)} \left(\int_{\Omega\times\Omega} c(x,y) \mathrm{d}\pi^\varepsilon(x,y) + \varepsilon \int_{\Omega\times\Omega} \log\left(\frac{\mathrm{d}\pi^\varepsilon(x,y)}{\mathrm{d}\rho(x)\,\mathrm{d}\sigma(y)} \right) \mathrm{d}\pi^\varepsilon(x,y) \right.$$
$$\left. - \varepsilon \int_{\Omega\times\Omega} \mathrm{d}\pi^\varepsilon(x,y) + \varepsilon \int_{\Omega\times\Omega} \mathrm{d}\rho(x)\,\mathrm{d}\sigma(y) \right). \quad (4.3)$$

The corresponding dual problem has the form

$$W_{2,\varepsilon}(\rho,\sigma)^2 = \max_{\psi^\varepsilon_\rho, \psi^\varepsilon_\sigma \in \mathcal{C}_b(\Omega)} \left(\int_\Omega \psi^\varepsilon_\rho(x) \mathrm{d}\rho(x) + \int_\Omega \psi^\varepsilon_\sigma(y) \mathrm{d}\sigma(y) \right.$$
$$\left. - \varepsilon \int_{\Omega\times\Omega} \exp\left(\frac{\psi^\varepsilon_\rho(x) + \psi^\varepsilon_\sigma(y) - c(x,y)}{\varepsilon} \right) \mathrm{d}\rho(x) \mathrm{d}\sigma(y) + \varepsilon \right). \quad (4.4)$$

Remark 4.2. *Note that we recover the constraint $\psi_\rho \oplus \psi_\sigma \leq c$ as $\varepsilon \to 0$ in Equation* (4.4)*.*

Remark 4.3. *In the entropic case, the form of the dual problem can be motivated by introducing the potentials as Lagrange multipliers of the form $\int_\Omega \psi_\rho \mathrm{d}\rho + \int_\Omega \psi_\sigma \mathrm{d}\sigma - \int_{\Omega\times\Omega}(\psi_\rho(x) + \psi_\sigma(y)) \mathrm{d}\pi(x,y)$ and calculating the stationarity condition for π.*

Remark 4.4. *Importantly, the derivative of $x \mapsto x \log \frac{x}{y} - x + y$ goes to $+\infty$ as $x \to 0^+$. This moves the optimizer of the primal problem to the interior of the admissible set. Most obvious is the following discrete case: the collocated form of Equation* (3.3) *with Dirac masses of equal weights $\hat\rho_i = \hat\sigma_j = 1/M \ \forall 1 \leq i,j \leq M$ reads*

$$\min_{\hat\pi \in \mathbb{R}^{M\times M}} \sum_{i,j=1}^{M} \hat c_{ij} \hat\pi_{ij} : \sum_i \hat\pi_{ij} = 1 \ \forall j = 1,\dots,M, \sum_j \hat\pi_{ij} = 1 \ \forall i = 1,\dots,M. \quad (4.5)$$

This is a linear minimization problem over bi-stochastic *matrices, which form a compact convex non-empty set. By the fundamental lemma of linear programming, the minimum will lie on the extreme points of this set, which are the permutation matrices. In contrast, the minimizer of*

$$\min_{\hat\pi \in \mathbb{R}^{M\times M}} \sum_{i,j=1}^{m} \left(\hat c_{ij} \hat\pi_{ij} + \varepsilon \hat\pi_{ij} \log \hat\pi_{ij} \right) : \sum_i \hat\pi_{ij} = 1 \ \forall j, \sum_j \hat\pi_{ij} = 1 \ \forall i \quad (4.6)$$

has strictly positive entries. Indeed, if $\hat{\pi}_{i^*j^*} = 0$ for some i^*, j^*, then for the family of plans $\hat{\pi}(t) := (1-t)\hat{\pi} + t/M^2$ for some $t > 0$, we have

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0^+} \sum_{i,j=1}^{m} (\hat{c}_{ij}\hat{\pi}(t)_{ij} + \varepsilon\hat{\pi}(t)_{ij}\log\hat{\pi}(t)_{ij}) = -\infty, \tag{4.7}$$

so $\hat{\pi}(t)\big|_{t=0}$ cannot be optimal.

Remark 4.5. *The entropic primal problem Equation (4.3) can be reformulated as*

$$W_{2,\varepsilon}(\rho,\sigma)^2 = \min_{\pi^\varepsilon\in\Pi(\rho,\sigma)} \varepsilon \int_{\Omega\times\Omega} \log\left(\frac{1}{k^\varepsilon(x,y)}\frac{\mathrm{d}\pi^\varepsilon(x,y)}{\mathrm{d}\rho(x)\,\mathrm{d}\sigma(y)}\right)\mathrm{d}\pi^\varepsilon(x,y), \tag{4.8}$$

where $k^\varepsilon(x,y) := \exp(-c(x,y)/\varepsilon)$ is called the Gibbs kernel. *This formulation is reminiscent of the* Schrödinger bridge problem*, see [86].*

A formal computation shows that the stationarity conditions for Equation (4.4) read

$$\exp\left(\frac{-\psi^\varepsilon_\sigma(y)}{\varepsilon}\right) = \int_\Omega \exp\left(\frac{\psi^\varepsilon_\rho(x) - c(x,y)}{\varepsilon}\right)\mathrm{d}\rho(x) \quad \sigma-\text{a. e.} \quad \text{and} \tag{4.9}$$

$$\exp\left(\frac{-\psi^\varepsilon_\rho(x)}{\varepsilon}\right) = \int_\Omega \exp\left(\frac{\psi^\varepsilon_\sigma(y) - c(x,y)}{\varepsilon}\right)\mathrm{d}\sigma(y) \quad \rho-\text{a. e..} \tag{4.10}$$

We call these two conditions the *Schrödinger equations.*

Remark 4.6. *At optimality, the double integral in Equation (4.4) evaluates to one due to Equation (4.9) and $W_{2,\varepsilon}(\rho,\sigma)^2$ is given by the sum of two weighted integrals of the potential functions.*

softmin and softmax

Taking the logarithm of Equation (4.9) defines the *softmin*, which replaces the *c-transform*:

Definition 4.2 (softmin)**.**

$$\begin{aligned}\psi^{c,\varepsilon}(y) &= -\varepsilon\log\int_\Omega \exp\left(\frac{\psi^\varepsilon(x) - c(x,y)}{\varepsilon}\right)\mathrm{d}\rho(x) \\ &=: \overset{\varepsilon}{\min_{x\sim\rho}}\{c(x,y) - \psi^\varepsilon(x)\}\end{aligned} \tag{4.11}$$

The parameter ε determines the strength of the regularization. A useful practical interpretation of ε is this: as the softmin operation is a Gaussian convolution, the entropic transport plan practically ignores features below the scale of $\sqrt{\varepsilon}$.

Proposition 4.1. *Assume that Ω is compact and denote by x^* the (not necessarily unique) $\arg\min_{x\in\operatorname{supp}\rho}(c(x,y) - \psi^\varepsilon(x))$ (recall that $x \mapsto c(x,y) - \psi^\varepsilon(x)$ is a bounded continuous function for all y). Then,*

$$\overset{\varepsilon}{\min_{x\sim\rho}}\{c(x,y) - \psi^\varepsilon(x)\} \overset{\varepsilon\to+\infty}{\longrightarrow} \int (c(x,y) - \psi^\varepsilon(x))\mathrm{d}\rho(x) \tag{4.12}$$

and

$$\overset{\varepsilon}{\min_{x\sim\rho}}\{c(x,y) - \psi^\varepsilon(x)\} \overset{\varepsilon\to 0}{\longrightarrow} \min_{x\in\operatorname{supp}\rho}(c(x,y) - \psi^\varepsilon(x)). \tag{4.13}$$

Proof. As all functions involved are smooth, the claim follows by Taylor expansion. Denote $f_y(x) := c(x,y) - \psi^\varepsilon(x)$. As $\varepsilon \to +\infty$,

$$\begin{aligned}\overset{\varepsilon}{\min_{x\sim\rho}}\{c(x,y)-\psi^\varepsilon(x)\} &= -\varepsilon\log\left(1-\varepsilon^{-1}\int f_y\mathrm{d}\rho + \frac{\varepsilon^{-2}}{2}\int f_y^2\mathrm{d}\rho + \mathcal{O}(\varepsilon^{-3})\right) && (4.14)\\ &= \int f_y\mathrm{d}\rho - \frac{\varepsilon^{-1}}{2}\int f_y^2\mathrm{d}\rho - \frac{\varepsilon^{-1}}{2}\left(\int f_y\mathrm{d}\rho\right)^2 + \mathcal{O}(\varepsilon^{-2}) && (4.15)\\ &\overset{\varepsilon\to+\infty}{\longrightarrow} \int(c(x,y)-\psi^\varepsilon(x))\mathrm{d}\rho(x). && (4.16)\end{aligned}$$

For $\varepsilon \to 0$,

$$\begin{aligned}\overset{\varepsilon}{\min_{x\sim\rho}}\{c(x,y)-\psi^\varepsilon(x)\} &= -\varepsilon\log\exp\left(\frac{-f_y(x^*)}{\varepsilon}\right)\\ &\quad+\varepsilon\log\int\exp\left(\frac{f_y(x^*)-f_y(x)}{\varepsilon}\right)\mathrm{d}\rho(x). && (4.17)\end{aligned}$$

By definition, the exponent appearing in the integral is negative and the integral is bounded by one. Therefore,

$$\overset{\varepsilon}{\min_{x\sim\rho}}\{c(x,y)-\psi^\varepsilon(x)\} \overset{\varepsilon\to0}{\longrightarrow} c(x^*,y)-\psi^\varepsilon(x^*). \tag{4.18}$$

□

Just as in the case with no regularization, the potentials are linked to convex functions, defined through an approximate maximum:

Proposition 4.2. *Let $c(x,y) = \frac{1}{2}|x-y|^2$. The function*

$$\begin{aligned}y \mapsto \frac{1}{2}|y|^2 - \psi^{c,\varepsilon}(y) &= \varepsilon\log\int_\Omega\exp\left(\frac{1}{\varepsilon}\left(x\cdot y - \frac{1}{2}|x|^2+\psi^\varepsilon(x)\right)\right)\mathrm{d}\rho(x) && (4.19)\\ &=: \overset{\varepsilon}{\max_{x\sim\rho}}\left\{x\cdot y - \left(\frac{|x|^2}{2}-\psi^\varepsilon(x)\right)\right\} && (4.20)\\ &=: \overset{\varepsilon}{\max_{x\sim\rho}}\{x\cdot y-\varphi^\varepsilon(x)\} && (4.21)\\ &=: \varphi^{*,\varepsilon}(y) && (4.22)\end{aligned}$$

is convex.

Proof. Evaluating the function at $y_t := ty_1 + (1-t)y_2 : 0 < t < 1$ gives

$$\begin{aligned}&\varepsilon\log\int_\Omega\exp\left(\frac{1}{\varepsilon}\left(x\cdot(ty_1+(1-t)y_2)-\varphi^\varepsilon(x)\right)\right)\mathrm{d}\rho(x)\\ &=\varepsilon\log\int_\Omega\left(\exp\left(\frac{1}{\varepsilon}\left(x\cdot y_1-\varphi^\varepsilon(x)\right)\right)\right)^t\left(\exp\left(\frac{1}{\varepsilon}\left(x\cdot y_2-\varphi^\varepsilon(x)\right)\right)\right)^{(1-t)}\mathrm{d}\rho(x).\end{aligned} \tag{4.23}$$

Applying Hölder's inequality with exponents $1/t, 1/(1-t)$ and using

$$\varepsilon\log\int(\dots)^t\mathrm{d}\rho \le \varepsilon\log\left(\int\dots\mathrm{d}\rho\right)^t = t\varepsilon\log\int\dots\mathrm{d}\rho \tag{4.24}$$

leads to

$$\varphi^{*,\varepsilon}(ty_1 + (1-t)y_2) \leq t\varphi^{*,\varepsilon}(y_1) + (1-t)\varphi^{*,\varepsilon}(y_2) \tag{4.25}$$

as claimed. □

Remark 4.7. *In fact, the convexity is strict as Hölder's inequality is an equality if and only if there exists a constant $C > 0$ such that $x \cdot (y_1 - y_2) = \varepsilon \log C$ for ρ-a.e. x, which would imply $y_1 = y_2$.*

Remark 4.8. *The limit cases $\varepsilon \to 0$: $\max^{\varepsilon}_{x\sim u} \to \max_{x\in \text{support}\,\rho}$ and $\varepsilon \to \infty$: $\max^{\varepsilon}_{x\sim\rho} \to \int_\Omega \mathrm{d}\rho$ are also convex.*

Remark 4.9. *The mapping $\psi \mapsto \psi^{c,\varepsilon}(y)$ is furthermore 1-Lipschitz for all y, i.e. $|\psi_1^{c,\varepsilon}(y) - \psi_2^{c,\varepsilon}(y)| \leq \|\psi_1 - \psi_2\|_{L^\infty}$ ([138], Proposition 17).*

Lastly, we improve the estimate for $\min^{\varepsilon}_{x\sim\rho}\{c(x,y) - \psi^\varepsilon(x)\}$ as $\varepsilon \to 0$.

Proposition 4.3. *Assume $\varphi^\varepsilon = \mathrm{id} - \psi^\varepsilon$ is strictly convex and let $\nabla\varphi^\varepsilon(\Omega) =: \Omega^\varepsilon$. For any $y \in \Omega^\varepsilon$, call x^* the point in Ω such that $\nabla\varphi^\varepsilon(x^*) = y$. Denote $\varphi^{*,\varepsilon}(y) = \max^{\varepsilon}_{x\sim\rho}\{x\cdot y - \varphi^\varepsilon(x)\}$. Assume that in a neighborhood around x^*, $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ is strictly positive and both φ^ε and ρ are smooth. Then,*

$$\begin{aligned}&\overset{\varepsilon}{\max_{x\sim\rho}}\{x\cdot y - \varphi^\varepsilon(x)\}\\ &\quad = x^*\cdot y - \varphi^\varepsilon(x^*) + \varepsilon \log\left((2\pi\varepsilon)^{d/2}\rho(x^*)(\det D^2\varphi^\varepsilon(x^*))^{-1/2}\left(1 + \sum_{j=1}^{\infty} a_j\varepsilon^j\right)\right)\end{aligned} \tag{4.26}$$

for coefficients $\{a_j\}_j$ depending on higher order derivatives of ρ and φ^ε at x^*.

Proof. The proposition is a direct application of Laplace's integral method, see Section 15.2 in [123]. □

Dual ascent algorithm

In practice, the entropic dual problem is solved by iteratively applying the softmin operation until the potentials fulfill Equation (4.9) and $\psi^\varepsilon_\rho = \psi^{c,\varepsilon}_\sigma$.

Proposition 4.4 ([138], Proposition 16)**.** *The sequence $\{(\psi^\varepsilon)^{(n)}, (\psi^{c,\varepsilon})^{(n)}\}_n$ obtained from*

$$(\psi^{c,\varepsilon})^{(n)}(y) := \overset{\varepsilon}{\min_{x\sim\rho}}\left\{c(x,y) - (\psi^\varepsilon)^{(n)}(x)\right\}, \tag{4.27}$$

$$(\psi^\varepsilon)^{(n+1)}(x) := \overset{\varepsilon}{\min_{y\sim\sigma}}\left\{c(x,y) - (\psi^{c,\varepsilon})^{(n)}(y)\right\} \tag{4.28}$$

converges in $(\mathcal{C}(\Omega), \|\cdot\|_{L^\infty})$ to the unique (up to a constant) potentials that maximize the dual problem Equation (4.4).

We refer to the reference for the proof. The strategy is much as in the unregularized case. The sequence $\{(\psi^\varepsilon)^{(n)}, (\psi^{c,\varepsilon})^{(n)}\}_n$ is both equi-bounded and equicontinuous, with modulus of continuity bounded by that of c (see Remark 4.9). This allows the extration of a converging subsequence in the sup-norm.

Furthermore, it holds that $\|\nabla^k\psi\|_\infty = \mathcal{O}(1 + \varepsilon^{1-k})$ [64].

Remark 4.10. *The solution to the primal problem* (4.3) *is given by*

$$\pi^\varepsilon = \rho \otimes \sigma \, \exp\left(\frac{1}{\varepsilon}\left(\psi_\rho^\varepsilon \oplus \psi_\sigma^\varepsilon - c\right)\right). \tag{4.29}$$

Remark 4.11. *The number of iterations needed to solve the entropic OT problem in practice go up dramatically as $\varepsilon \to 0$. In particular, in cases where the solution for $\varepsilon = 0$ is a smooth map, the error after the lth iteration is of order $(1-\varepsilon)^l$ ([106], Remark 4.15).*

In general, the number of iterations is expected to be of order $\|c\|_{L^\infty}/\varepsilon$ [83, 24, 119, 84].

For moderately small values of ε (when compared to the characteristic scale of the cost function), this is not yet very restrictive. Beyond that, however, simulated annealing and multiscale methods become necessary (c.f. [60], Section 3.3.3 and [106], Section 4.2).

Convergence as $\varepsilon \to 0$

The properties of $W_{2,\varepsilon}$, including its convergence as ε goes to zero, have been studied extensively. We give here the result from [38] for the quadratic cost. The proof therein, based on Γ-convergence, is recommended to the reader as it is rather short and fully self-contained. For more general results, we refer to [85, 100, 20].

Theorem 4.1 (Convergence of entropic optimal transport ([38], Theorem 2.7))**.** *Let $\Omega \subset \mathbb{R}^d$ bounded and $\rho, \sigma \in \mathcal{P}(\Omega)$ have finite entropy, i.e. they are absolutely continuos with respect to the Lebesgue measure and $\int \log\rho \,\mathrm{d}\rho < +\infty$, $\int \log\sigma \,\mathrm{d}\sigma < +\infty$. Let $\{\varepsilon_k\}_k$ be a non-negative sequence converging to zero. Then,*

$$\lim_{k\to+\infty} W_{2,\varepsilon_k}(\rho,\sigma) = W_2(\rho,\sigma). \tag{4.30}$$

Furthermore, the optimal transport plans π_{ε_k} in $W_{2,\varepsilon_k}(\rho,\sigma)$ converge narrowly to the optimal transport plan π in $W_{2,\varepsilon_k}(\rho,\sigma)$.

Remark 4.12. *The potentials of the entropic dual problem also converge to the potentials of the unregularized dual problem (in the $L^1(\Omega,\rho)$ and $L^1(\Omega,\sigma)$ norm, respectively), see [100], Theorem 1.1.*

The entropic transport map

The entropic OT problem does not admit a transport map as a solution, as the transport plan is necessarily supported on the entirety of $\rho \otimes \sigma$. It is a natural question what the map $x \mapsto x - \nabla\psi_\rho^\varepsilon(x)$ corresponds to. From the stationarity condition Equation (4.9), we find, for $c(x,y) = \frac{1}{2}|x-y|^2$,

$$\begin{aligned}
\nabla\psi_\rho^\varepsilon(x) &= \frac{\int (x-y)\exp\left(\left(\psi_\sigma^\varepsilon(y) - c(x,y)\right)/\varepsilon\right)\mathrm{d}\sigma(y)}{\int \exp\left(\left(\psi_\sigma^\varepsilon(y) - c(x,y)\right)/\varepsilon\right)\mathrm{d}\sigma(y)} && (4.31)\\
&= x - \frac{\int y\exp\left(\left(\psi_\sigma^\varepsilon(y) - c(x,y)\right)/\varepsilon\right)\mathrm{d}\sigma(y)}{\int \exp\left(\left(\psi_\sigma^\varepsilon(y) - c(x,y)\right)/\varepsilon\right)\mathrm{d}\sigma(y)} && (4.32)\\
&=: x - T_{\rho\to\sigma}^\varepsilon(x) && (4.33)
\end{aligned}$$

Definition 4.3 (Entropic transport map). *We call*

$$T^{\varepsilon}_{\rho\to\sigma} = \mathrm{id} - \nabla\psi^{\varepsilon}_{\rho} \tag{4.34}$$

the entropic transport map *between ρ and σ.*

It must be stressed that $T^{\varepsilon}_{\sharp}\rho \neq \sigma$ in general. Nevertheless, the map has appealing properties: It is defined for all $y \in \Omega$ (not only ρ - almost everywhere) and converges to the transport map of the unregularized problem as $\varepsilon \to 0$.

The entropic transport map can also be interpreted as an extension of the expected value $x \mapsto \mathbb{E}_{\pi^{\varepsilon}}[Y|X = x]$ from support(ρ) to the entire domain. It is also referred to as the *barycentric mapping* [60] or *barycenter projection* of the transport plan π_{ε} [107] as it has the form of a weighted mean with normalized weights.

In [107], the convergence of an entropic transport map obtained from M samples from ρ and σ, respectively, to the unregularized optimal transport map between ρ and σ is investigated. Using an optimal choice of $\varepsilon(M)$, it is established that the empirical entropic transport map obtained from the samples converges to the true transport map $T_{\rho\to\sigma}$ with approximate rate $M^{-1/d}$ under certain regularity assumptions on the densities and $\varphi : \nabla\varphi = T_{\rho\to\sigma}$. There are numerous other results regarding the convergence of such regularized empirical transport costs, plans, and maps [6, 64, 20].

A central property of the transport maps was that we were able to define their inverse through the c-transform transform of the corresponding potentials. Proposition 4.3 suggests that this is the case in the entropic case as well. We can show that it holds up to order ε:

$$\begin{aligned} \nabla\varphi^{*,\varepsilon}(y) &= \frac{\int x \exp\left(\left(x\cdot y - \varphi^*(x)\right)/\varepsilon\right) \mathrm{d}\rho(x)}{\int \exp\left(\left(x\cdot y - \varphi^*(x)\right)/\varepsilon\right) \mathrm{d}\rho(x)} && (4.35)\\ &= \frac{x^*(1+\mathcal{O}(\varepsilon))}{1+\mathcal{O}(\varepsilon)}, && (4.36) \end{aligned}$$

since the factors of the leading order from Proposition 4.3 are identical in the nominator and denominator.

4.2 The Sinkhorn algorithm

Let us consider the discrete case of $\rho(x) = \sum_{i=1}^{M} \hat{\rho}\delta_{x_i}$ and $\sigma(y) = \sum_{j=1}^{M} \hat{\sigma}\delta_{y_j}$. We can without loss of generality assume that $\hat{\rho}_i$ and $\hat{\sigma}_j$ are strictly positive for all i, j. Indeed, points without mass would not play any role in the optimization.

Log-domain formulation

The iterative updates of the potentials in Equation (4.27) take the form

$$(\hat{\psi}^{c,\varepsilon})^{(n)}_j \leftarrow -\varepsilon \log \sum_{i=1}^{M} \exp\left(\frac{(\hat{\psi}^{\varepsilon}_i)^{(n)} - \hat{c}_{ij}}{\varepsilon} + \log\hat{\rho}_i\right) \qquad \forall 1 \le j \le M \tag{4.37}$$

$$(\hat{\psi}^{\varepsilon})^{(n+1)}_i \leftarrow -\varepsilon \log \sum_{j=1}^{M} \exp\left(\frac{(\hat{\psi}^{c,\varepsilon})^{(n)}_j - \hat{c}_{ij}}{\varepsilon} + \log\hat{\sigma}_j\right) \qquad \forall 1 \le j \le M. \tag{4.38}$$

We define the *logsumexp* function as

$$\mathrm{LSE} : \mathbb{R}^M \to \mathbb{R} : \mathtt{f} \mapsto \log \sum_{i=1}^{M} \exp \mathtt{f}_i. \tag{4.39}$$

The LSE function is a commonly used tool in computer science, in particular data analysis and machine learning. A naive implementation can cause numerical overflow problems, since $\sum_i \exp \mathtt{f}_i$ can take very large values.

This problem can be mitigated by letting $\mathtt{f}^* = \max_i \mathtt{f}_i$ and using the identity $\mathrm{LSE}(\mathtt{f}) = \mathtt{f}^* + \mathrm{LSE}(\mathtt{f} - \mathtt{f}^*\mathtt{1})$, where now all elements of the sum are bounded by one. The symbol $\mathtt{1}$ denotes a vector of ones.

An even better approach, used in the package `LogExpFunctions.jl`[1] and described in [99], is to calculate the maximum element in a streaming manner, updating it when necessary. This avoids the one additional loop over $\mathtt{f}$.

Matrix-scaling formulation

The operations carried out in the iterative updates Equation (4.27) can be expressed as matrix-vector operations. We introduce the *scaling factors* $\mathtt{a}, \mathtt{b} \in \mathbb{R}^M$ with entries

$$\mathtt{a}_i := \exp(\hat{\psi}^\varepsilon_i/\varepsilon) \text{ and } \mathtt{b}_j := \exp(\hat{\psi}^{c,\varepsilon}_j/\varepsilon) \quad \forall i, j = 1, \ldots, M \tag{4.40}$$

and collocated Gibbs kernel $\mathtt{K} \in \mathbb{R}^{M \times M}$ with entries

$$\mathtt{K}_{ij} := \exp(-\hat{c}_{ij}/\varepsilon) \quad \forall i, j = 1, \ldots, M. \tag{4.41}$$

With respect to these quantities, the iterates on the dual problem can be written in the form of matrix-vector products and element-wise operations on vectors entirely, see Algorithm 4. The operations $.\!\!\leftarrow, .*, ./$ denote element-wise assignment, multiplication, and division, respectively. We write $\mathtt{C}$ for the $M \times M$ matrix with entries $\hat{c}_{ij}$ for all i, j.

Algorithm 4 Sinkhorn's algorithm

```
function SINKHORN($\hat{\rho}, \hat{\sigma}, \mathtt{c}, \varepsilon, \mathrm{tol}$)
    $\mathtt{a}, \mathtt{b} \,.\!\!\leftarrow \mathtt{1}$
    $\mathtt{K} \,.\!\!\leftarrow \exp.(-\mathtt{C}/\varepsilon)$
    while $\|\hat{\rho} - \hat{\rho} .* \mathtt{a} .* \mathtt{K}(\mathtt{b} .* \hat{\sigma})\|_1 > \mathrm{tol}$ do    ▷ $l^1$ error of the marginal condition
        $\mathtt{a} \,.\!\!\leftarrow \mathtt{1} ./ \mathtt{K}(\mathtt{b} .* \hat{\sigma})$
        $\mathtt{b} \,.\!\!\leftarrow \mathtt{1} ./ \mathtt{K}(\mathtt{a} .* \hat{\rho})$
    end while
    return $\varepsilon \log.\mathtt{a}, \varepsilon \log.\mathtt{b}$    ▷ The Kantorovich potentials
end function
```

In this form, the values of ε that can be used are restricted by issues of numerical stability and overflow, as elements of $\mathtt{K}$ can become extremely small and elements of `a, b` can become extremely large. On the other hand, the algorithm is easy to implement and very fast.

[1] https://github.com/JuliaStats/LogExpFunctions.jl

Nomenclature

Algorithm 4 is named after the author of [125], where the convergence of the method in the discrete case is shown. In particular, this work is concerned with finding the representation of a matrix with strictly positive entries as the product of a diagonal matrix, a bi-stochastic matrix, and another diagonal matrix. The procedure itself is much older and known in different fields as *iterative proportional fitting procedure*, *RAS*[2] method, *gravity method*, *iterative Bregman projections*, or *softassign*. Remark 4.5 in [106] and Section 3.3.1 of [60] give historical overviews and more context.

Stabilization

In [41, 119], a strategy of stabilizing the algorithm without moving it entirely into the log-domain is given. It relies on the fact that $\psi^\varepsilon \oplus \psi^{c,\varepsilon} - c$ remains bounded and close to zero throughout the iterations. A redundant parametrization of the form $\mathtt{a} = \mathtt{a}' \mathbin{.*} \exp((\hat{\psi}^\varepsilon)'_i/\varepsilon)$ (and analogously for $\mathtt{b}$) is used, where extreme values of $\mathtt{a}'$ are absorbed by $(\hat{\psi}^\varepsilon)'$ every few iterations. The entries of the stabilized kernel $\mathtt{K}'_{ij} := \exp((\hat{\psi}^\varepsilon)'_i + \hat{\psi}^{c,\varepsilon})'_j - \mathtt{C}_{ij})/\varepsilon)$ stay under control and $\mathtt{Kb} = \exp.(\hat{\psi}^\varepsilon)'/\varepsilon) \mathbin{.*} \mathtt{K}'(\mathtt{b}' \mathbin{.*} \hat{\sigma})$.

Separable kernels

Naive implementations of both the matrix-vector products from Algorithm 4 and the LSE evaluation in (4.37) are of $\mathcal{O}(N^2)$ complexity due to nested loops over i and j. Note that M scales exponentially with the spatial dimension of the problem when it is discretized on a grid. However, in the special case of $c(x,y) = |x-y|^2$ we are working with, we can do better, as pointed out in [126]. Note that in dimension d, the cost is separable in d terms along each dimension: $|x-y|^2 = |x^1-y^1|^2 + \cdots + |x^d-y^d|^2$. Now assume the points x_i are sampled on a regular tensor grid and therefore can be indexed as $x_{i_1,\dots,i_d} : 1 \le i_1,\dots,i_d \le M^{1/d}$. Note that $x^l_{i_1,\dots,i_d}$ can be denoted with $x^l_{i_l}$, as only the lth coordinate changes when varying the indices $i_1,\dots,i_d$. Let $\mathtt{C}^l_{ij} := |x^l_{i_l} - y^l_{j_l}|^2$ and $\mathtt{K}^l = \exp.(-\mathtt{C}^l/\varepsilon)$ for $l = 1,\dots,d$. In this case,

$$(\mathtt{Ka})_j = \sum_{1 \le i \le M} \mathtt{K}_{ij}\mathtt{a}_i \quad \forall j = 1,\dots,M \tag{4.42}$$

is equal to

$$\begin{aligned}(\mathtt{Ka})_{j_1,\dots,j_d} &= \sum_{1 \le i_1,\dots,i_d \le M^{1/d}} \mathtt{K}^1_{i_1 j_1} \cdots \mathtt{K}^d_{i_d j_d} \mathtt{a}_{i_1,\dots,i_d} \\ &= \sum_{1 \le i_1 \le M^{1/d}} \mathtt{K}^1_{i_1 j_1} \sum_{1 \le i_2 \le M^{1/d}} \mathtt{K}^2_{i_2 j_2} \cdots \sum_{1 \le i_d \le M^{1/d}} \mathtt{K}^d_{i_d j_d} \mathtt{a}_{i_1,\dots,i_d} \end{aligned} \tag{4.43}$$

for all $j_1,\dots,j_d \in 1,\dots,M$. As a result, instead of computing one large matrix-vector product of complexity M^2, we are computing d tensor contractions of complexity $M^{1+1/d}$ each.

An analogous trick can be applied in the log-domain as well.

[2] Interestingly, the origin of this name appears to be unknown [33]

Annealing strategies

As one decreases ε, the number of iterations needed to reach convergence in Algorithm 4 increases approximately as $\|c\|_{L^\infty}/\varepsilon$. The strategy of *$\varepsilon$-scaling*, where ε is initialized with a rather large value initially and then gradually decreased, was introduced for the auction and Sinkhorn algorithm already in [21] (citing an earlier unpublished paper from 1987) and [79]. It has proven very efficient in practice, decreasing the iterations needed until convergence to $\mathcal{O}(\log(\|c\|_{L^\infty}/\varepsilon))$ [60, 119].

The heuristic is that the steps when ascending the regularized dual problem are of size ε.

When employing ε-scaling, large steps are taken initially to find an approximate coupling between ρ and σ, which is then gradually refined as ε is reduced.

In practice, the scaling can be done by starting at, e.g. $\varepsilon^{(0)} = \|c\|_{L^\infty}$ and then multiplying it by some scaling factor $s \in (0,1)$ in every iteration, such that $\varepsilon^{(l)} = s^l \varepsilon^{(0)}$. Therefore, if we set $\varepsilon^{(l)} =: \varepsilon$, this implies $l = \log(\|c\|_{L^\infty}/\varepsilon)/\log(1/s)$ iterations. Just as in line-search methods, the choice of s constitutes a trade-off between speed and safety.

Further modifications

A number of other modifications to the Sinkhorn algorithm have been proposed and shown to increase its speed. Among them are multiscale methods [93, 60, 119], low-rank approximations of the kernel matrix `K` [4], methods where only select entries of `a,b` are updated in each iteration in a greedy manner [5], regularized nonlinear acceleration methods [132], and momentum methods [117]. In [60], the author also advocates the use of averaged updates of the dual potentials, i.e.

$$(\psi^{c,\varepsilon})^{(n+1)}(y) \leftarrow \frac{1}{2}(\psi^{c,\varepsilon})^{(n)}(y) + \frac{1}{2} \overset{\varepsilon}{\min_{x\sim\rho}} \left\{ c(x,y) - (\psi^{\varepsilon})^{(n)}(x) \right\}, \tag{4.44}$$

$$(\psi^{\varepsilon})^{(n+1)}(x) \leftarrow \frac{1}{2}(\psi^{\varepsilon})^{(n)}(x) + \frac{1}{2} \overset{\varepsilon}{\min_{y\sim\sigma}} \left\{ c(x,y) - (\psi^{c,\varepsilon})^{(n)}(y) \right\}. \tag{4.45}$$

This modification ensures symmetry of the computed entropic transport distance at every iteration. Especially when the two input measures ρ and σ are very close to each other, significant speed-up of convergence is reported.

Another interesting extension is [126], where $c(x,y) = d(x,y)^2$, the squared geodesic distance on a compact Riemannian manifold. The authors note that in the Euclidean case, the Gibbs kernel

$$k(x,y) = \exp\left(-\frac{|x-y|^2}{2\varepsilon}\right)$$

corresponds to the heat kernel at time $t = \varepsilon/2$. Hence, convolution of a function against k is equivalent to solving the heat equation $\partial_t u = \Delta u$ until time $t = 2\varepsilon$ with that function as an initial condition. The rigorous justification of this idea is given by Varadhan's formula [89].

4.3 Sinkhorn divergences

Despite the widespread applications of $W_{2,\varepsilon}$ and the efficient algorithms available to compute it, there is a substantial downside. The function $W_{2,\varepsilon} : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}_{\geq 0}$

does not define a distance.

This is most obvious in the limit where $\varepsilon \to \infty$: the minimization in Equation (4.3) will select $\pi = \rho \otimes \sigma$, so $W_{p,\varepsilon}(\rho,\sigma)^2 \overset{\varepsilon\to+\infty}{\longrightarrow} \int c(x,y)\mathrm{d}\rho(x)\mathrm{d}\sigma(y)$. In this limit, $\min_\sigma W_{p,\infty}(\rho,\sigma)^2 \neq \rho$. The measure that achieves $\min_\sigma W_{p,\infty}(\rho,\sigma)^2 = \min_\sigma \int |x-y|^2\mathrm{d}\rho(x)\mathrm{d}\sigma(y)$ is a Dirac measure centered at the mean of ρ, denoted δ_{m_ρ}. To see this, notice that

$$\begin{aligned}\int |x-y|^2\mathrm{d}\rho(x)\mathrm{d}\sigma(y) &= \int |x|^2\mathrm{d}\rho(x) + \int |y|^2\mathrm{d}\sigma(y) - 2m_\sigma\cdot m_\rho && (4.46)\\ &\geq \int |x|^2\mathrm{d}\rho(x) + m_\sigma^2 - 2m_\sigma\cdot m_\rho && (4.47)\\ &\geq \int |x|^2\mathrm{d}\rho(x) - m_\sigma^2 && (4.48)\\ &= W_{p,\infty}(\rho,\delta_{m_\rho})^2. && (4.49)\end{aligned}$$

As a result, using $W_{p,\varepsilon}$ as loss function in optimization is bound to lead to unsatisfactory results - unless one wants to explicitly use the *entropic bias* as a denoising method [111].

Debiasing

In [109], the authors introduce a modification of the entropic Wasserstein distance that does vanish when comparing the same measure.

Definition 4.4 (Sinkhorn divergence). *Given two measures $\rho,\mu \in \mathcal{P}(\Omega)$ and $\varepsilon > 0$, their* (debiased) Sinkhorn divergence *S_ε is given by*

$$S_\varepsilon(\rho,\sigma) := W_{2,\varepsilon}(\rho,\sigma)^2 - \frac{1}{2}W_{2,\varepsilon}(\rho,\rho)^2 - \frac{1}{2}W_{2,\varepsilon}(\sigma,\sigma)^2. \tag{4.50}$$

Remark 4.13. *Introducing the optimal transport potential $\psi^\varepsilon_{\rho\leftrightarrow\rho}$ and $\psi^\varepsilon_{\sigma\leftrightarrow\sigma}$ for $W_{2,\varepsilon}(\rho,\rho)^2$ and $W_{2,\varepsilon}(\sigma,\sigma)^2$, respectively,*

$$\frac{1}{2}S_\varepsilon(\rho,\sigma) = \int_\Omega \left(\psi^\varepsilon_{\rho\to\sigma} - \psi^\varepsilon_{\rho\leftrightarrow\rho}\right)\mathrm{d}\rho + \int_\Omega \left(\psi^\varepsilon_{\sigma\to\rho} - \psi^\varepsilon_{\sigma\leftrightarrow\sigma}\right)\mathrm{d}\sigma. \tag{4.51}$$

Theorem 4.2 (Properties of the Sinkhold divergence ([60], Theorem 3.1)). *When $\Omega \subset \mathbb{R}^d$ is compact and $\rho,\mu \in \mathcal{P}(\Omega)$ with bounded support, then*

$$\begin{aligned} 0 = S_\varepsilon(\rho,\rho) &\leq S_\varepsilon(\rho,\sigma) && (4.52)\\ S_\varepsilon(\rho,\sigma) = 0 &\Leftrightarrow \rho = \sigma && (4.53)\\ \rho_n \rightharpoonup \rho &\Leftrightarrow S_\varepsilon(\rho_n,\rho) \to 0 && (4.54)\end{aligned}$$

for any $\varepsilon > 0$.

Proof and discussion of Theorem 4.2 can be found in appendix A of [60]. The Sinkhorn divergence can be computed with the same iterative algorithm as the entropic Wasserstein distance, extended by computation of the debiasing potentials corresponding to the $W_{2,\varepsilon}(\rho,\rho)^2$ and $W_{2,\varepsilon}(\sigma,\sigma)^2$ terms, which are always solved using a symmetric update rule. We present it in Algorithm 5 in its matrix-scaling form.

Algorithm 5 Debiased Sinkhorn algorithm

```
function SINKHORN($\hat{\rho}, \hat{\sigma}, \mathtt{c}, \varepsilon, \mathrm{tol}$)
    $\mathtt{a}, \mathtt{b}, \mathtt{d}_\rho, \mathtt{d}_\sigma \; .\!\leftarrow \mathtt{1}$
    $\mathtt{K} \; .\!\leftarrow \exp.(-\mathtt{C}/\varepsilon)$
    while $\|\hat{\rho} - \hat{\rho} \; .\!* \; \mathtt{a} \; .\!* \; \mathtt{K}(\mathtt{b} \; .\!* \; \hat{\sigma})\|_1 > \mathrm{tol}$ do
        $\mathtt{a} \; .\!\leftarrow \mathtt{1} \; ./ \; \mathtt{K}(\mathtt{b} \; .\!* \; \hat{\sigma})$
        $\mathtt{b} \; .\!\leftarrow \mathtt{1} \; ./ \; \mathtt{K}(\mathtt{a} \; .\!* \; \hat{\rho})$
        $\mathtt{d}_\rho \; .\!\leftarrow (\mathtt{d}_\rho \; ./ \; \mathtt{K}(\mathtt{d}_\rho \; .\!* \; \hat{\rho})) \; .^{1/2}$
        $\mathtt{d}_\sigma \; .\!\leftarrow (\mathtt{d}_\sigma \; ./ \; \mathtt{K}(\mathtt{d}_\sigma \; .\!* \; \hat{\sigma})) \; .^{1/2}$
    end while
    return $\varepsilon \log.\mathtt{a} - \varepsilon \log.\mathtt{d}_\rho, \varepsilon \log.\mathtt{b} - \varepsilon \log.\mathtt{d}_\sigma$    ▷ Debiased transport potentials
end function
```

The formulation in the log-domain is analogous and can be found, for example, in [60], Algorithm 3.4.

The analogue to the entropic transport map from Definition 4.3 is

$$T^{\varepsilon}_{S,\rho\to\sigma} : x \mapsto x - \nabla\psi^{\varepsilon}_{\rho\to\sigma}(x) - \nabla\psi^{\varepsilon}_{\rho\leftrightarrow\rho}(x). \tag{4.55}$$

Remark 4.14. *The modifications and improvements to the Sinkhorn algorithm can be applied to the debiased variant as well. In particular, this includes separation of kernels and ε-scaling and from Section 4.2 and Section 4.2. The convergence criterion can also be modified if one wants to avoid computing the l^1 error of the marginal condition. For example, a relative tolerance on the size of the updates of the potentials can be used.*

4.4 Methods without regularization

To conclude this chapter, we want to remark on a number of other computational methods that have been developed to solve optimal transport problems. These are not the focus of the present work and therefore their presentation will be rather short. That is not to say that they are inferior to the schemes presented so far, on the contrary, some of these methods have very strong links to PDE theory which could possibly be used in our application case.

Direct method for the two-dimensional case

In [96], the authors consider the problem of optimal transport on a rectangular domain $\subset \mathbb{R}^2$. They approximate transport maps $T_{\rho\to\sigma}$ by $T_{\bar{\rho}\to\sigma} \circ T_{\rho\to\bar{\rho}}$, akin to the Monge embedding approach. When $\bar{\rho}$ is chosen constant and equal to $|\Omega|^{-1}$, the Monge-Ampére equation for ψ, the optimal transport potential corresponding to the transport from ρ to $\bar{\rho}$, simplifies to $\det(\mathrm{Id} - D^2\psi) = \rho$.

In two dimensions, this equation satisfies

$$\det(\mathrm{Id} - D^2\psi) = 1 - \Delta\psi + \det D^2\psi = \rho. \tag{4.56}$$

Equation (4.56) is an equality, and should not be confused with the approximation $\det(\mathrm{Id} + \epsilon D^2 f) \approx 1 + \epsilon\Delta f$ for small $\epsilon > 0$. It allows a very fast solution of the Monge-Ampére equation by an iterative scheme:

The approach relies on solving the Poisson equation $-\Delta\psi^{(n)} = \rho - 1 + \det D^2\psi^{(n-1)}$ where $\det D^2\psi^{(n-1)}$ is treated as a source term. After $\psi^{(n)}$ is computed, $\det D^2\psi^{(n)}$ is calculated and used as a source term in the next iteration.

The boundary conditions in the special case of a rectangular domain are of Neumann type: $\nabla\psi \cdot \hat{n} = 0$ on $\partial\Omega$ with $\hat{n}$ the unit outward normal vector. The map $T_{\bar{\rho}\to\sigma}$ is obtained by inverting $T_{\sigma\to\bar{\rho}}$. The latter can be computed using the same iterative method that was used to obtain $T_{\rho\to\bar{\rho}}$.

The authors prove convergence of this scheme using a quasi-Newton method using a special conjugate gradient descent. By pre-factoring the Laplace operator, the computation is extremely fast, depending on the problem size up to factors 50 faster than dynamical and semi-discrete schemes, which we discuss later.

Dynamical schemes

The method we describe next goes back to [16]. Overviews of the method, its variants and applications can be found in [115], Chapter 6, as well as [15]. More details and proofs of convergence can be found in [71, 81]. Two examples of implementation are [104], a finite difference version, and [98], based on mixed finite elements. The presentation we give largely follows [71].

Through the Bennamou-Brenier formula, it is possible to reformulate the optimal transport problem to a minimization of the kinetic energy of the flow transporting a measure ρ_0 to ρ_1, see Section 3.6. The evolution of the density throughout this transport will be given by a curve $t \mapsto \rho(t,\cdot)$. Let us denote by v the (time-dependent) vector field generating the transporting flow and introduce a new variable, the momentum $M := \rho v$. In these variables, we seek

$$\min_{\rho,M}\left(\int_0^1\int_\Omega \frac{|M(t,x)^2|}{2\rho(t,x)}\mathrm{d}x\mathrm{d}t : \partial_t\rho + \nabla_x M = 0 \text{ and either } \rho > 0 \text{ or } \rho = M = 0\right). \tag{4.57}$$

The boundary conditions (in the space-time domain) are given by $\rho(0,\cdot) = \rho_0$, $\rho(1,\cdot) = \rho_1$, and no-flux boundary conditions on $\partial\Omega$ for $t \in (0,1)$. Introducing a Lagrange multiplier χ, the problem can be written as the following saddle-point problem:

$$\inf_{\rho,M}\sup_{\chi}\left(\int_{(0,1)\times\Omega}\frac{|M|^2}{2\rho} - \int_{(0,1)\times\Omega}(\partial_t\chi\rho - M\cdot\nabla_x\chi) + \int_\Omega(\chi(0,\cdot)\rho_0 - \chi(1,\cdot)\rho_1)\right). \tag{4.58}$$

What is missing so far is the constraint $\rho > 0$. The crucial trick is to write the kinetic energy density function $\mathcal{K}(\rho, M) := |M|^2/2\rho$ as the Legendre transform of its Legendre transform

$$\mathcal{K}^*(a,B) = \sup_{\rho,M}\left(a\rho + B\cdot M - \frac{|M|^2}{2\rho}\right) = \begin{cases} 0 & \text{if } a \le -|B|^2/2 \\ +\infty & \text{else.}\end{cases} \tag{4.59}$$

By considering $\mathcal{K}^{**}(\rho, M)$, one guarantees the necessary constraints on ρ, since $\mathcal{K}^{**}(\rho, M) = +\infty$ when $\rho > 0$. All together, the problem can be written as

$$\inf_{\chi,(a,B)}\sup_{\rho,M} L(\chi,(a,B),(\rho,M)), \tag{4.60}$$

where the Lagrangian $L(\chi, (a, B), (\rho, M))$ consists of the sum of three terms: the characteristic function of the (convex) set $\{(a, B) : a \leq -|B|^2/2\}$, the second term in Equation (4.58), and the bi-linear term

$$\langle(\rho, M), (\partial_t \chi, \nabla_x \chi) - (a, B)\rangle_{L^2((0,1)\times\Omega)}. \tag{4.61}$$

This saddle point problem on the space of measures (M is a measure valued in $\mathbb{R}^d$), can be solved with first-order optimization methods from convex analysis and the *proximal splitting algorithm* in particular. Note that the non-linearity in M is replaced by the presence of the indicator function. We refer to the references given at the start of this section for further details.

Solving the dynamical optimization problem takes place in $\mathbb{R}^{d+1}$ due to the introduction of a time variable. This of course increases the computational cost of the method. One of the advantages is that additional dynamics and constraints, such as upper or lower bounds on ρ or forcing and interaction terms can be added into the action functional.

Semi-discrete OT

This case deals with the transport between a continuous density σ and a density $\rho = \sum_{i=1}^{M} \hat{\rho}_i \delta_{x_i}$ localized on a number of Dirac measures. The dual problem in this case turns into

$$\max_{\psi \in \mathcal{C}_b(\Omega)} \left(\sum_{j=1}^{M} \hat{\rho}_j \psi(x_j) + \int \psi^c \mathrm{d}\sigma \right). \tag{4.62}$$

All points y such that $\psi(x_j) + \psi^c(y) = c(x_j, y)$ will be transported to x_j, since

$$\psi^c(y) = \inf_x (c(x, y) - \psi(x)) \leq c(x_k, y) - \psi(x_k) \quad \forall k = 1, \ldots, M. \tag{4.63}$$

When c is the quadratic cost, and if all Dirac measures have the same weight, every point of mass is transported to the closest Dirac measure. The result are *Voronoi cells.*

Through the additional factor $\psi(x_j)$, every cell around δ_{x_j}, which we will denote V_j, gets an additional parameter which can increase or decrease its volume relative to this case. V_j are called *Laguerre cells* and, as Voronoi cells, are convex polyhedra. At optimality, it holds that

$$\hat{\rho}_j = \int_{V_j} \mathrm{d}\sigma \quad \forall j = 1, \ldots, M. \tag{4.64}$$

The computation of power cells is for example implemented in the `Geogram`[3] library, see also [93, 88, 78, 87]. These methods allow the solution of semi-discrete OT problems with up to 10^6 Dirac measures.

We mention these methods because a number of interesting papers use this approach to derive Lagrangian schemes to solve equations from fluid dynamics, in particular the incompressible Euler equation, see [63, 95] and Appendix B.

[3] https://github.com/BrunoLevy/geogram

Chapter 5

Reduced models in $\mathcal{P}(\Omega)$

The theory of optimal transport is interesting in its own right as a field of pure mathematics, for example because of the connection to the Monge-Ampère equation and a number of other known PDEs (see Appendix B). However, its properties also make it an attractive tool for several applications in data science, as we motivate now.

5.1 Motivation

Example 5.1. *Assume that we are given a target measure $\sigma \in \mathcal{P}(\Omega)$ and a family of measures $\{\rho(\mu) : \mu \in \mathcal{A}\} \subset \mathcal{P}(\Omega)$ that are parametrized through μ. Suppose we want to find an optimal value μ^* such that $\sigma \approx \rho(\mu^*)$, i.e.*

$$\mu^* = \underset{\mu \in \mathcal{A}}{\arg\min}\, \mathrm{Loss}(\rho(\mu), \sigma). \tag{5.1}$$

The setting of Example 5.1 is that of a *measure-fitting* or *registration* problem. It arises in applications of image processing, computational anatomy, and also in the construction of *generative adversarial networks* (GANs) [68]. A possible solution strategy is then a gradient descent method along the lines of

$$\mu^{(n+1)} \leftarrow \mu^{(n)} - t\nabla_\mu \mathrm{Loss}(\rho(\mu^{(n)}), \sigma) \tag{5.2}$$

with some step size t.

Loss functions

The success of such a method hinges on the correct choice of loss function, which needs to be differentiable and capture the discrepancy between measures in a meaningful way. Note that these measures might be challenging to handle. When $\Omega \subset \mathbb{R}^2$, the features might be supported on curves or even points, hence the measures give mass to small sets.

Even if the measures admit a smooth density, L^p type norms might fail completely. As long as the supports of $\rho(\mu)$ and σ do not overlap, their L^1 distance is simply equal to two and the gradient is zero. The problem can persist even if the measures are strictly positive as we have seen in Example 2.6.

Other notions of discrepancy that are commonly used such as the KL divergence share this problem. In particular, the KL divergence $\mathrm{KL}(\rho|\sigma)$ requires ρ to be absolutely continuous with respect to σ. The large class of *kernel norms* can also struggle with the problem. Norms based on kernels that are not heavy-tailed are blind to features lie far apart and at the same time, kernels that are too smooth will not be able to resolve small-scale differences. This is evident in the case of the very well known Gaussian kernel $(x, y) \mapsto k_\epsilon(x, y) := \exp(-|x - y|^2/2\varepsilon)$, which is for all intents and purposes equal to zero once $|x-y| > 3\sqrt{\varepsilon}$. These phenomena are known in the literature as *vanishing gradients* and *electric shielding.* We refer to Section 3.2 of [60] for an excellent summary of the topic.

The optimal transportation distance looks like an attractive choice for this application. It can handle very general measures, gives meaningful distances even when their support is far apart (Proposition 3.5), and, as we will see, has a computable gradient.

Remark 5.1. *Another comment is in order regarding the terminology* Wasserstein-GAN *(W-GAN) [9]. In generative neural networks, the distance between measures is measured in a weak sense by testing $\rho(\mu)-\sigma$ with a family of* discriminator *functions g_Θ, represented by a neural network and parametrized by a set of weights Θ. Taking the most strict discriminator as the loss function, one arrives at*

$$\mathrm{Loss}(\rho(\mu), \sigma) := \max_{\Theta} \int g_\Theta \mathrm{d}(\rho(\mu) - \sigma). \tag{5.3}$$

It is clear that without any constraints on g_Θ, the discriminators can be too critical: for example, allowing step functions in the set of $\{g_\Theta\}_\Theta$, the discrepancy between any measures with disjoint support can be made arbitrarily large. In W-GANs, the set of discriminators is (loosely) enforced to be 1-Lipschitz. The problem

$$\max_{\psi:\ \textit{1-Lipschitz}} \int \psi \mathrm{d}(\rho(\mu) - \sigma) \tag{5.4}$$

has the form of the dual optimal transport problem with cost $c(x, y) = |x-y|$. Indeed, the constraint in the dual problem $\psi(x) + \psi^c(y) \leq |x - y|$ is precisely the 1-Lipschitz condition, since $\psi^c = -\psi$ in this case ([115], Proposition 3.1).

It is important to note, however, that the optimization over g_Θ is not the same as an optimization over all 1-Lipschitz functions. The latter is a very large space, while the elements of $\{g_\Theta\}_\Theta$ should be designed to encode a problem-specific notion of similarity that is unlikely to coincide with the W_1 distance.

The second argument for the use of the optimal transport is that the displacement interpolation is a natural choice in several applications where a standard weighted mean fails. We have already seen this in Figure 3.2, where the former lead to a physical advection-diffusion-like transport while the latter corresponds to nonphysical teleportation of mass. In [73], it is shown that the displacement interpolation between a measure ρ_0 and ρ_t coincides with the solution of a PDE with initial condition ρ_0 on the interval $(0, t)$ up to a re-scaling of time and multiplicative constants for the heat equation $\partial_t \rho = \Delta \rho$, the non-linear diffusion equation $\partial_t \rho = \Delta \rho^m$ with $1 < m \in \mathbb{N}$, and the Riemann problem of the Sod shock tube problem.

5.2 Differentiability of W_2

When we use the W_2 distance as a loss function, the vast majority of optimization algorithms used in measure fitting applications require us to compute its gradient with respect to the input measures.

Recall the dual formulation of the optimal transport distance, Equation (3.5):

$$W_2(\rho,\sigma)^2 = \max_{\psi_\rho,\psi_\sigma\in\mathcal{C}(\Omega)} \left\{ \int_\Omega \psi_\rho \mathrm{d}\rho + \int_\Omega \psi_\sigma \mathrm{d}\sigma \ : \ \psi_\rho(x)+\psi_\sigma(y) \leq c(x,y) \right\}.$$

A formal derivative of $W_2(\rho,\sigma)$ therefore will return the optimal transport potential ψ_ρ that maximizes the objective plus additional terms that follow from the implicit dependence of ψ_ρ and ψ_σ (which, as we know, equals ψ_ρ^c) on ρ.

However, one might hope that these additional terms in fact do not contribute, since the variation of the objective with respect to the potentials should vanish when evaluated at optimal potentials. In this case, the derivative of $W_2(\rho,\sigma)$ with respect to ρ would be just ψ_ρ.

This statement is sometimes called the *envelope theorem* in optimization: If we assume that $x_\mu^* := \arg\min_x F(\mu,x)$ for a set of μs is the family of solutions to a smooth, convex optimization problem in $\mathbb{R}^d$ parametrized by μ, then necessarily $\nabla_x F(\mu,x_\mu^*)=0\ \forall\mu$ and

$$\frac{\mathrm{d}}{\mathrm{d}\mu}F(\mu,x_\mu^*) = \frac{\partial}{\partial\mu}F(\mu,x_\mu^*) + \frac{\partial x_\mu^*}{\partial\mu}\cdot\underbrace{\nabla_x F(\mu,x_\mu^*)}_{=0} = \frac{\partial}{\partial\mu}F(\mu,x_\mu^*). \tag{5.5}$$

In the case of $W_2(\rho,\sigma)$, μ corresponds to ρ while x_μ corresponds to the pair of optimal transport potentials (ψ_ρ,ψ_ρ^c).

Indeed, the assumed result can be made rigorous.

Theorem 5.1 (Subdifferential of the W_2 distance [116], Proposition 4.8)**.** *Let $\rho,\sigma\in\mathcal{P}(\Omega)$. The function $\rho\to W_2(\rho,\sigma)^2$ is convex and its subdifferential at ρ_0 is the set of optimal transport potentials*

$$\left\{\psi\in\mathcal{C}(\Omega) : \frac{1}{2}W_2(\rho_0,\sigma)^2 = \int\psi\mathrm{d}\rho_0 + \int\psi^c\mathrm{d}\sigma\right\}. \tag{5.6}$$

If there is one (up to an additive constant) unique c-concave optimal transport potential ψ_{ρ_0}, then

$$\frac{\delta W_2(\cdot,\sigma)^2}{\delta\rho}(\rho_0) = 2\psi_{\rho_0}. \tag{5.7}$$

A proof for the case of compact Ω can be found in [115], Proposition 7.17. We also have the following, related result:

Theorem 5.2 (Chain rule for the W_2 derivative [140], Theorem 8.13)**.** *Let $\Omega\subset\mathbb{R}^d$ and $(\epsilon,\epsilon)\ni t\mapsto\rho_t$ be an absolutely continuous curve through $\mathcal{P}_{\mathrm{ac}}(\Omega)$, i.e. there exists an L^2 integrable $t\mapsto c(t)$ such that $W_2(\rho_s,\rho_t)\leq\int_s^t c(\tau)\mathrm{d}\tau\ \forall s,t$. Furthermore, let $\sigma\in\mathcal{P}_{\mathrm{ac}}(\Omega)$ and assume $t\mapsto\rho_t$ satisfies*

$$\partial_t\rho_t + \nabla\cdot(v_t\rho_t) = 0 \tag{5.8}$$

for a globally bounded $(x,t) \mapsto v_t(x) \in \mathcal{C}^1(\Omega, (-\epsilon, \epsilon); \mathbb{R}^d)$. Then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} W_2(\rho_t, \sigma)^2 \bigg|_{t=0} &= 2 \int \nabla\psi \cdot v_0 \mathrm{d}\rho_0 \\ &= 2 \int (\mathrm{id} - \nabla\varphi) \cdot v_0 \mathrm{d}\rho_0 \\ &= -2 \int (\mathrm{id} - \nabla\varphi^*) \cdot (v_0 \circ \nabla\varphi^*) \mathrm{d}\sigma, \end{aligned} \tag{5.9}$$

where $\nabla\varphi$ is the optimal transport map between ρ_0 and σ.

We refer to the reference for a proof. Note that the result agrees with what another formal computation yields:

$$\frac{\mathrm{d}}{\mathrm{d}t} W_2(\rho_t, \sigma)^2 = \int \frac{\delta W_2(\rho_t, \sigma)^2}{\delta\rho} \partial_t \mathrm{d}\rho_t = 2 \int \psi_{\rho_t} \partial_t \mathrm{d}\rho_t = 2 \int \nabla\psi_{\rho_t} \cdot v_t \mathrm{d}\rho_t, \tag{5.10}$$

where we used the continuity equation in the last step. The last line in Equation (5.9) follows from a change of variables using $\nabla\varphi^*_\sharp \sigma = \rho_0$ and the identity $\nabla\varphi \circ \nabla\varphi^* = \mathrm{id}$.

Remark 5.2. *The results from Theorem 5.1 and Theorem 5.2 extend to the case with entropic regularization, i.e.*

$$\frac{1}{2} \frac{\delta W_{2,\varepsilon}(\rho_0, \sigma)^2}{\delta\rho} = \psi^\varepsilon_{\rho_0} \tag{5.11}$$

and

$$\frac{1}{2} \frac{\delta S_\varepsilon(\rho_0, \sigma)^2}{\delta\rho} = \psi^\varepsilon_{\rho_0 \to \sigma} - \psi^\varepsilon_{\rho_0 \leftrightarrow \rho_0}. \tag{5.12}$$

Sketch of proof. Denote by ψ^ε_t the optimal transport potentials for the transport between ρ_t and σ and by $\psi^{c,\varepsilon}_t$ its c-transform. Furthermore, recall that at optimality, $\frac{1}{2} W_{2,\varepsilon}(\rho_t, \sigma)^2 = \int \psi^\varepsilon_t \mathrm{d}\rho_t + \int \psi^{c,\varepsilon}_t \mathrm{d}\sigma$. Now, use the fact that ψ^ε_0 is suboptimal for $W_{2,\varepsilon}(\rho_t, \sigma)$, hence

$$\frac{1}{2} W_{2,\varepsilon}(\rho_t, \sigma)^2 \geq \int \psi^\varepsilon_0 \mathrm{d}\rho_t + \int \psi^{c,\varepsilon}_0 \mathrm{d}\sigma - \varepsilon \int \exp\left(\frac{\psi^\varepsilon_0 \oplus \psi^{c,\varepsilon}_0 - c}{\varepsilon} \right) \mathrm{d}\rho_t \mathrm{d}\sigma + \varepsilon. \tag{5.13}$$

As $(\psi^\varepsilon_0, \psi^{c,\varepsilon}_0)$ are optimal for (ρ_0, σ), $\int \exp\left((\psi^\varepsilon_0(x) + \psi^{c,\varepsilon}_0(y) - c(x,y))/\varepsilon \right) \mathrm{d}\sigma(y) = 1$ ρ_0-a.e., which by assumption is ρ_t-a.e., so the last two terms cancel. We conclude that

$$\frac{1}{2} W_{2,\varepsilon}(\rho_t, \sigma)^2 - \frac{1}{2} W_{2,\varepsilon}(\rho_0, \sigma)^2 \geq \int \psi^\varepsilon_0 \mathrm{d}(\rho_t - \rho_0). \tag{5.14}$$

Repeating the same considerations with the roles of ρ_t and ρ_0 reversed yields

$$\frac{1}{2} W_{2,\varepsilon}(\rho_t, \sigma)^2 - \frac{1}{2} W_{2,\varepsilon}(\rho_0, \sigma)^2 \leq \int \psi^\varepsilon_t \mathrm{d}(\rho_t - \rho_0). \tag{5.15}$$

By the stability of optimality, as $\rho_t \rightharpoonup \rho$, $\psi^\varepsilon_t \to \psi^\varepsilon_0$ uniformly, so that

$$\frac{1}{2} \frac{W_{2,\varepsilon}(\rho_t, \sigma)^2 - W_{2,\varepsilon}(\rho_0, \sigma)^2}{t} \overset{t \to 0^+}{\longrightarrow} \int \psi^\varepsilon_0 \partial_t \mathrm{d}\rho_0. \tag{5.16}$$

□

5.3 Wasserstein barycenters

The first variational problem in $(\mathcal{P}(\Omega), W_2)$ we consider is the following: Given a set of measures $\{\rho_i\}_{i=1}^m$, which element of $P(\Omega)$ minimizes the sum of W_2 distances to each element of this set?

An element that solves this problem is an average of the input elements $\{\rho_i\}_i$ in the sense of the optimal transport distance.

In fact, the displacement interpolation $\rho_t := ((1-t)\mathrm{id} + tT_{\rho_0 \to \rho_1})_\sharp \rho_0$ is a special case for $m = 2$. The displacement interpolation defines a weighted convex combination of elements in $(\mathcal{P}(\Omega), W_2)$. The extension to arbitrary $m \in \mathbb{N}$ we denote the *optimal transport barycenter* (the names *Wasserstein barycenter*, *Karcher mean*, or *Fréchet mean* are also used).

Definition 5.1 (Optimal transport barycenter). *Let $\{\rho_i\}_{i=1}^m$ a set of probability measures on Ω and $\{\omega_i\}_{i=1}^m$ a set of positive real numbers that sum to one, i.e. $\{\omega_i\}_{i=1}^m \in \Sigma_m$ We call*

$$\inf_{\sigma \in \mathcal{P}(\Omega)} \frac{1}{2} \sum_{i=1}^m \omega_i W_2(\rho_i, \sigma)^2 \tag{5.17}$$

the optimal transport barycenter problem. A solution of this problem will be called the barycenter of $\{\rho_i\}_i$ with weights $\{\omega_i\}_i$ and denoted $W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i)$.

For a number of applications, the optimal transport barycenter provides a smooth transition between input elements for applications in shape interpolation. In particular, we refer to the examples in [126].

5.3.1 Existence, uniqueness, and properties

We recall the following results without proof:

Proposition 5.1 ([1], Propositions 2.3, 3.8 and Theorem 5.1). *The optimal transport barycenter problem admits the dual formulation*

$$\sup \left\{ \sum_{i=1}^m \int \inf_{y \in \Omega} \left(\frac{\omega_i}{2} |x-y|^2 - \psi_i(y) \right) \mathrm{d}\rho_i(x) : \psi_1, \ldots, \psi_m \in \mathcal{C}_b(\Omega), \sum_{i=1}^m \psi_i = 0 \right\}. \tag{5.18}$$

Both primal and dual problems admit solutions and their values coincide. As long as one of $\{\rho_i\}_i$ is absolutely continuous, Equation (5.17) *has a unique solution. If all of $\{\rho_i\}_i$ are absolutely continuous, the transport potentials ψ_i between ρ_i and $W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i)$ satisfy*

$$\sum_{i=1}^m \omega_i \nabla \psi_i^c = 0 \tag{5.19}$$

and this condition is sufficient for optimality. Furthermore, if ρ_1 has a density bounded from above, then

$$\|W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i)\|_{L^\infty} \le \omega_i^{-d} \|\rho_1\|_{L^\infty}. \tag{5.20}$$

Remark 5.3. *In [1], the notation ψ is used for the convex function whose gradient is the transport map, denoted φ in this work. In this variable, the condition reads $\sum_j \omega_j \nabla \varphi_j^* = \mathrm{id}$. The two are equivalent as $\nabla \varphi_j^* = \mathrm{id} - \nabla \psi_j^c$ and $\sum_j \omega_j = 1$.*

Note that we can obtain the condition Equation (5.19) from a variational argument of the form

$$0 \overset{!}{=} \frac{\mathrm{d}}{\mathrm{d}\epsilon}\frac{1}{2}\sum_i \omega_i W_2(\rho_i, \sigma^\epsilon)^2 = \sum_i \omega_i \int \nabla\psi_i^c \cdot v \,\mathrm{d}\sigma \quad \forall v, \tag{5.21}$$

using the results from Section 5.2.

Similar to the factorization of translations from Proposition 3.5, we have the following result:

Proposition 5.2. *Let $\{\rho_i\}_{i=1}^m \subset \mathcal{P}(\Omega)$ with at least one element $\in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and $\{\omega_i\}_{i=1}^m \in \Sigma_m$. Then,*

$$\int x \,\mathrm{d}(W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_{i=1}^m))(x) = \sum_{i=1}^m \omega_i \int x \,\mathrm{d}\rho_i(x). \tag{5.22}$$

Proof. Let $\sigma := W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i) \in \mathcal{P}_{\mathrm{ac}}(\Omega)$. Note that $(\mathrm{id} - \nabla\psi_j^c)_\sharp\sigma = \rho_j \ \forall j = 1, \dots, m$. Hence,

$$\sum_i \omega_i \int x\mathrm{d}\rho_i(x) = \sum_i \omega_i \int x\mathrm{d}((\mathrm{id} - \nabla\psi_i^c)_\sharp\sigma)(x) \tag{5.23}$$

$$= \int y\mathrm{d}\sigma(y) - \int \left(\sum_i \omega_i \nabla\psi_j^c(y)\right) \mathrm{d}\sigma(y) \tag{5.24}$$

and the last term vanishes by Equation (5.19). □

Lastly, we recall the following result that characterizes the properties of the map $\{\rho_i, \omega_i\}_i \to W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i)$ that is crucial for the encoding approaches discussed in Section 5.3.5.

Proposition 5.3 ([51], Lemma 2.1)**.** *Let $\{\rho_i\}_{i=1}^m \subset \mathcal{P}(\Omega)$ and $\{\omega_i\}_{i=1}^m \in \Sigma_m$. The map $(\{\rho_i\}_i, \{\omega_i\}_i) \to W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i)$ is continuous in its second argument and lower-semicontinuous in its first argument with respect to narrow convergence.*

Furthermore, if Ω is compact and $\{\rho_i\}_i$ has at least one element $\in \mathcal{P}_{\mathrm{ac}}(\Omega)$, then the map $\{\omega_i\}_i \to W_2\mathrm{Bar}(\{\rho_i, \omega_i\}_i)$ with $\{\rho_i\}_i$ fixed is differentiable in Σ_m. In this case, the set

$$\left\{\sigma \in \mathcal{P}(\Omega) : \exists \{\omega_i\}_i \in \Sigma_m : \sigma = W_2\mathrm{Bar}(\{\omega_i; \rho_i\}_i)\right\} \tag{5.25}$$

of all possible barycenters of $\{\rho_i\}_i$ is weakly sequentially compact in $(\mathcal{P}(\Omega), W_2)$.

5.3.2 Entropic regularization and bias

An *entropic optimal transport barycenter* can be defined by replacing W_2 in Equation (5.17) by $W_{2,\varepsilon}$. However, as we have already seen in Section 4.3, optimization problems with respect to the entropic transport cost are subject to the phenomenon of entropic bias: Let $\pi, \varsigma \in \mathcal{P}(\Omega \times \Omega)$ and

$$\mathrm{KL}(\pi|\varsigma) := \int_{\Omega\times\Omega} \left(\log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\varsigma}\right)\mathrm{d}\pi - \mathrm{d}\pi + \mathrm{d}\varsigma\right). \tag{5.26}$$

We observe that

$$W_{2,\varepsilon}(\rho,\sigma)^2 = \min_{\pi^\varepsilon\in\Pi(\rho,\sigma)} \left(\int_{\Omega\times\Omega} c(x,y)\mathrm{d}\pi^\varepsilon(x,y) + \varepsilon\mathrm{KL}(\pi^\varepsilon|\rho\otimes\sigma) \right) \tag{5.27}$$

and

$$\mathrm{KL}(\pi|\varsigma_1) = \mathrm{KL}(\pi|\varsigma_2) + \mathrm{KL}(\varsigma_2|\varsigma_1)\ \forall\varsigma_1,\varsigma_2 \in \mathcal{P}(\Omega\times\Omega). \tag{5.28}$$

Let $W_{2,\varepsilon}^{\varsigma}(\rho,\sigma)^2 := \min_{\pi^\varepsilon\in\Pi(\rho,\sigma)}(\int_{\Omega\times\Omega} c(x,y)\mathrm{d}\pi^\varepsilon(x,y) + \varepsilon\mathrm{KL}(\pi^\varepsilon|\varsigma))$. The choice $\varsigma = \rho\otimes\sigma$ in the definition of $W_{2,\varepsilon}(\rho,\sigma)$ (no superscript) is natural, as π^ε is guaranteed to be absolutely continuous with respect to this product measure. However, many other choices (e.g. one that is constant on $\mathrm{supp}(\rho\otimes\sigma)$) are also possible. Replacing ς_1 by ς_2 changes the value of $W_{2,\varepsilon}^{\varsigma_1}$ by the constant $\mathrm{KL}(\varsigma_2|\varsigma_1)$.

Smoothing and shrinking

When computing entropic optimal transport barycenters, a different choice of ς leads to

$$W_{2,\varepsilon}\mathrm{Bar}^{\varsigma}(\{\omega_i;\rho_i\}_i) = \operatorname*{arg\,min}_{\sigma\in\mathcal{P}(\Omega)} \sum_{k=1}^{m} \omega_k \left(W_{2,\varepsilon}(\rho_k,\sigma)^2 + \varepsilon\mathrm{KL}(\varsigma|\rho_k\otimes\sigma) \right). \tag{5.29}$$

Selecting a constant ς leads to entropic smoothing. To minimize the second term and reduce the discrepancy between $\rho_k\otimes\sigma$ and the constant ς, the barycenter is blurred. This effect is discussed in [45, 126]. In [74], it is shown that when all input measures are Gaussian $\rho_k = \mathcal{N}(\mu_k, \mathrm{var})$, the entropic barycenter using the Lebesgue measure in KL will be a Gaussian, namely $\mathcal{N}(\sum_k \omega_k\mu_k, \mathrm{var}+\varepsilon)$. The choice of the product measure leads to $\mathcal{N}(\sum_k \omega_k\mu_k, \mathrm{var}-\varepsilon)$ for $\mathrm{var} > \varepsilon$ and $\delta_{\sum_k \omega_k\mu_k}$ otherwise.

This smoothing (resp. shrinking) bias can be seen as a feature of the method, as in [111] it is shown that the entropic shrinking corresponds to a maximum-likelihood deconvolution technique. A smoothing of the barycenter can also be beneficial in applications, especially as it can translate to more regular transport maps.

Sinkhorn divergence barycenters

Alternatively, one can remove the effect of the choice in KL by replacing $W_{2,\varepsilon}$ with S_ε: It is straightforward to show that the value of S_ε no longer depends on ς, but only on $\mathrm{KL}(\pi^\varepsilon_{\rho,\sigma}|\pi^\varepsilon_{\rho,\rho})$ and $\mathrm{KL}(\pi^\varepsilon_{\rho,\sigma}|\pi^\varepsilon_{\sigma,\sigma})$. For the example of Gaussians $\rho_k = \mathcal{N}(\mu_k, \mathrm{var})$, the S_ε barycenter is $\mathcal{N}(\sum_k \omega_k\mu_k, \mathrm{var})\ \forall\varepsilon > 0$ ([74], Theorem 3) and coincides with the true W_2 barycenter.

Definition 5.2 (Entropic optimal transport barycenters). *We call*

$$\min_{\sigma\in\mathcal{P}(\Omega)} \frac{1}{2}\sum_{i=1}^{n} \omega_i W_{2,\varepsilon}(\rho_i,\sigma)^2 \text{ and } \min_{\sigma\in\mathcal{P}(\Omega)} \frac{1}{2}\sum_{i=1}^{n} \omega_i S_\varepsilon(\rho_i,\sigma)^2 \tag{5.30}$$

the entropic barycenter problem and debiased entropic barycenter problem, respectively.

5.3.3 Computation

In [74], following [17], a modified Sinkhorn algorithm is presented that can compute the debiased optimal transport barycenter for a set of measures, represented by a list of vectors via collocation as before. We repeat the method in Algorithm 6.

We denote the multiple input densities by $\rho_{(i)}$, and the corresponding scaling factors by $\mathtt{a}_{(i)}, \mathtt{b}_{(i)} \in \mathbb{R}^M$ to emphasize the difference to individual components of these vectors.

Algorithm 6 Debiased Sinkhorn barycenter algorithm

```
 1: function BARYCENTER($\{\hat\rho_{(i)}\}_{i=1}^m, \{\omega_i\}_{i=1}^m, \mathtt{c}, \varepsilon, \mathrm{tol}$)
 2:     for $i = 1, \dots, m$ do
 3:         $\mathtt{a}_{(i)}, \mathtt{b}_{(i)} \mathrel{.\!\!\leftarrow} \mathtt{1}$
 4:     end for
 5:     $\mathtt{d} \mathrel{.\!\!\leftarrow} \mathtt{1}$
 6:     $\mathtt{K} \mathrel{.\!\!\leftarrow} \exp.(-\mathtt{C}/\varepsilon)$
 7:     while $\max_i \|\hat\rho_{(i)} - \mathtt{a}_{(i)} \,.\!* \mathtt{K}\mathtt{b}_{(i)}\|_1 > \mathrm{tol}$ do
 8:         for $i = 1, \dots, m$ do
 9:             $\mathtt{a}_{(i)} \mathrel{.\!\!\leftarrow} \hat\rho_{(i)} \,./\, \mathtt{K}\mathtt{b}_{(i)}$
10:         end for
11:         $\hat\sigma \mathrel{.\!\!\leftarrow} \mathtt{d} \,.\!*\, . \prod_{i=1}^m (\mathtt{K}\mathtt{a}_{(i)}).^{\omega_i}$
12:         for $i = 1, \dots, m$ do
13:             $\mathtt{b}_{(i)} \mathrel{.\!\!\leftarrow} \hat\sigma \,./\, \mathtt{K}\mathtt{a}_{(i)}$
14:         end for
15:         $\mathtt{d} \mathrel{.\!\!\leftarrow} (\mathtt{d} \,./\, \mathtt{K}\mathtt{d}_\sigma)\,.^{1/2}$
16:     end while
17:     return $\hat\sigma$                ▷ The debiased barycenter $S_\varepsilon\mathrm{Bar}(\{\omega_i; \rho_i\}_i)$
18: end function
```

When compared to Algorithm 5, the scaling factors $\mathtt{a}_{(i)}, \mathtt{b}_{(i)}$, and $\mathtt{d}$ are defined differently, namely as

$$\mathtt{a}_{(i)} \,.\!*\, \hat\rho_{(i)} := \exp.(\hat\psi^\varepsilon_{(i)}/\varepsilon) \tag{5.31}$$

$$\mathtt{b}_{(i)} \,.\!*\, \hat\sigma := \exp.(\hat\psi^{c,\varepsilon}_{(i)}/\varepsilon) \quad \text{and} \tag{5.32}$$

$$\mathtt{d} \,.\!*\, \hat\sigma := \exp.(\hat\psi^\varepsilon_{\sigma\leftrightarrow\sigma}/\varepsilon), \tag{5.33}$$

where $\psi^\varepsilon_{(i)} = \psi^\varepsilon_{\rho_i\to\sigma}$. With this definition, Line 11 in Algorithm 6 corresponds to the stationarity condition

$$\sum_i \omega_i(\psi_i^{c,\varepsilon} - \psi^\varepsilon_{\sigma\leftrightarrow\sigma}) = \sum_i \omega_i\psi_i^{c,\varepsilon} - \psi^\varepsilon_{\sigma\leftrightarrow\sigma} = 0. \tag{5.34}$$

5.3.4 Monge embedding barycenters

In $\mathbb{R}^d$, the Euclidean barycenter of a collection of points $\{x_i\}_i$ has the simple closed form $\mathrm{Bar}(\{\omega_i; x_i\}_i) = \sum_i \omega_i x_i$. On $T_\rho\mathcal{P}(\Omega)$, we can define an analogous object:

Definition 5.3 (Monge embedding barycenter)**.** *Let $\bar\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega), \{\sigma_i\}_i \subset \mathcal{P}(\Omega)$, and $\{\omega_i\}_i \in \Sigma_m$. We define the* Monge embedding barycenter *of $\{\sigma_i\}_i$ with weights $\{\omega_i\}_i$ as*

$$\mathrm{ME}_{\bar\rho}\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m) := \left(\sum_{i=1}^m \omega_i T_{\bar\rho\to\sigma_i}\right)_\sharp \bar\rho = \sharp_{\bar\rho} \circ \left(\sum_{i=1}^m \omega_i \mathrm{ME}_{\bar\rho}(\sigma_i)\right). \tag{5.35}$$

Proposition 5.4. *Let $\bar\rho, \{\sigma_i\}_i, \{\omega_i\}_i$ as in Definition 5.3. If $(\bar\rho, \sigma_i, T_{\rho\to\sigma_i})$ are compatible for all $i = 1, \dots, m$, then*

$$\mathrm{ME}_{\bar\rho}\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m) = W_2\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m). \tag{5.36}$$

Sketch of proof. In the compatible case,

$$W_2\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m) = \arg\min_\nu \frac{1}{2}\sum_i \omega_i \|T_{\bar\rho\to\nu} - T_{\bar\rho\to\sigma_i}\|^2_{L^2(\bar\rho)} \tag{5.37}$$

$$= \arg\min_\nu \frac{1}{2}\sum_i \omega_i \|\mathrm{ME}_{\bar\rho}(\nu) - \mathrm{ME}_{\bar\rho}(\sigma_i)\|^2_{L^2(\bar\rho)} \tag{5.38}$$

For $\mathrm{ME}_{\bar\rho}(\nu)$, this is a quadratic problem with stationarity condition

$$\sum_i \omega_i \left(\mathrm{ME}_{\bar\rho} \circ W_2\mathrm{Bar}(\{\omega_j;\sigma_j\}_{j=1}^m) - \mathrm{ME}_{\bar\rho}(\sigma_i)\right) = 0. \tag{5.39}$$

Therefore, $\mathrm{ME}_{\bar\rho} \circ W_2\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m) = \sum_{i=1}^m \omega_i \, \mathrm{ME}_{\bar\rho}(\sigma_i)$ and

$$W_2\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m) = \sharp_{\bar\rho}\circ\mathrm{ME}_{\bar\rho}\circ W_2\mathrm{Bar}(\{\omega_i;\sigma_i\}_{i=1}^m) = \sharp_{\bar\rho}\circ\left(\sum_{i=1}^m \omega_i \, \mathrm{ME}_{\bar\rho}(\sigma_i)\right). \tag{5.40}$$

□

Note that the condition $\{\omega_i\}_i \in \Sigma_m$ guarantees that $\sum_{i=1}^m \omega_i \, \mathrm{ME}_{\bar\rho}(\sigma_i)$ is itself again an optimal transport map, as it can be written as the gradient of a convex combination of convex functions.

5.3.5 Barycenter encoding

Optimal transport barycenters allow us to take convex combinations of measures on $(\mathcal{P}(\Omega), W_2)$. We now consider the encoding problem of approximating one measure $\rho \in \mathcal{P}(\Omega)$ as a convex combination of a set $\{\sigma_i\} \subset \mathcal{P}(\Omega)$.

Given this set of probability measures $\{\sigma_i\}_i$, recall the set of all possible barycenters of $\{\sigma_i\}_i$ from Equation (5.25):

$$\{\rho \in \mathcal{P}(\Omega) : \exists\{\omega_i\}_i \in \Sigma_m : \rho = W_2\mathrm{Bar}(\{\omega_i;\sigma_i\}_i)\}.$$

This set is similar in principle to the space spanned by a set of basis functions $\{\zeta_i\}_i$, however it is of course not a linear space. One can see it as a non-linear *dictionary learning* approach, where a high dimensional feature is encoded using a set of *atoms* that form the dictionary and a list of *codes* that relates data points and codes.

Dictionary approximations

More precisely, let us assume that the data is given in form of a large $N \times n_s$ matrix $\mathbb{S}$ whose columns are the data points. The atoms $\{a_i\}_{i=1}^m$ are a set of m vectors in $\mathbb{R}^N$ and the jth code w_j is represented as a vector in $\mathbb{R}^m$. In the case where the reconstruction of data from codes is linear, we have

$$\mathbb{S}_{ij} \approx \sum_{k=1}^{m} (a_i)_m (w_j)_m \quad \forall i, j. \tag{5.41}$$

If we want to minimize the approximation error of $\mathbb{S}$ in the Frobenius norm, we find ourselves back in the setting of the Schmidt-Eckart-Young theorem and the discussion from Section 2.2. Different norms such as l^1 or l^0 can be used to promote sparsity of the representation [2].

In the non-linear case, codes and atoms can be related by a non-linear reconstruction function. This is the setting of Equation (5.25), where this reconstruction is the mapping $\{\omega_i; \sigma_i\}_i \mapsto W_2\mathrm{Bar}(\{\omega_i; \sigma_i\}_i)$. The *encoding* step that assigns an element in $\mathcal{P}$ a weight vector is defined as

Definition 5.4 (Barycenter encoding). *Given a set of measures $\{\sigma_i\}_i \subset \mathcal{P}(\Omega)$, we define the* barycenter encoding *or* barycenter coordinates *of $\rho \in \mathcal{P}(\Omega)$ as*

$$\{\omega_i^\rho\}_i \in \operatorname*{arg\,min}_{\{\omega_i\}_i \in \Sigma_m} \mathrm{Loss}\left(\rho, W_2\mathrm{Bar}(\{\omega_i; \sigma_i\}_i)\right). \tag{5.42}$$

The loss function $\mathrm{Loss} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ *is free to choose, one can for example use* $\mathrm{Loss}(\rho, \sigma) = W_2(\rho, \sigma)^2$.

Recall that Proposition 5.3 gave sufficient conditions for the map

$$\{\omega_i\}_i \mapsto W_2\mathrm{Bar}(\{\omega_i; \sigma_i\}_i) \tag{5.43}$$

to be differentiable.

Optimization using iterative scaling methods

In [27], the authors apply the barycentric encoding to a number of problems from computer graphics. In these examples, the atoms selected by the user and represent a range of images or shapes. The optimal transport barycenter is then used as a non-linear interpolation method. Barycenters are computed using entropic regularization and a number of loss functions (L^1, L^2, KL, and W_2) are investigated. The fitting problem itself is solved using standard quasi-Newton methods. While the optimization proved robust, the authors note that the optimal weights are often sparse. Since sparse weight vectors correspond to faces of the simplex Σ_m, this could hint towards the fact that the addition of another atom could improve the approximation quality.

Linear characterization

A different approach is proposed in [143]. Therein, the authors use the following fact: if $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ can be expressed as the barycenter of $\{\sigma_i\}_i \subset \mathcal{P}_{\mathrm{ac}}(\Omega)$, all bounded from above and below, for some weights $\{\omega_i^\rho\}_i \in \Sigma_m$, then necessarily

$$\left\| \frac{\delta}{\delta\rho} \sum_{i=1}^{m} \omega_i^\rho W_2(\sigma_i, \rho)^2 \right\|_{L^2(\rho)}^2 = \sum_{i,j=1}^{m} \omega_i^\rho \omega_j^\rho \langle \nabla\psi_{\rho\to\sigma_i}, \nabla\psi_{\rho\to\sigma_j} \rangle_{L^2(\rho)} \overset{!}{=} 0. \tag{5.44}$$

Note that the matrix $\{\langle \nabla\psi_{\rho\to\sigma_i}, \nabla\psi_{\rho\to\sigma_j} \rangle_{L^2(\rho)}\}_{1\le i,j\le m}$ must not have full rank for the stationarity condition to be fulfilled, as solutions correspond to eigenvectors of zero eigenvalues. If the rank is smaller than $m-1$, there might be several solutions, showing redundancy in the set of atoms. Note that these conditions are not sufficient for solutions to exist, as the weight vectors are also constrained to lie in Σ_m.

Furthermore, it cannot be expected in practice that Equation (5.44) admits a solution, i.e. that ρ can be exactly expressed as a weighted barycenter of $\{\rho_i\}_i$.

The (approximate) barycenter coordinates are then defined as those weights that minimize $\{\omega_i\}_i \mapsto \|\frac{\delta}{\delta\rho}\sum_{i=1}^m \omega_i W_2(\sigma_i,\rho)^2\|_{L^2(\rho)}^2$. This approach avoids the solution of a non-linear and in general non-convex optimization problem. However, unless the minimum is exactly zero, solving the quadratic problem from Equation (5.44) is only equivalent to the exact barycenter encoding problem if $(\rho, \sigma_i, T_{\rho\to\sigma_i})$ are compatible for all $i = 1, \ldots, m$, as we have seen in Section 5.3.4. In general, the minimization on $\mathcal{P}(\Omega)$ gives a different result than that on $T_\rho\mathcal{P}(\Omega)$.

5.3.6 Wasserstein dictionary learning

In [92, 117], the authors do not work with a given set of atoms, but instead determine them by optimization.

Definition 5.5 (Wasserstein dictionary learning). *Given a training set* $\{\rho_i\}_{i=1}^{n_s} \subset \mathcal{P}(\Omega)$ *and* $m \in \mathbb{N}$, *the* Wasserstein dictionary learning *problem determines atoms* $\{\sigma_j^*\}_{j=1^m}$ *and weights* $\omega_j^*(\rho_i)\}_j \in \Sigma_m$ *by solving*

$$\min_{\substack{\{\omega_j(\rho_i)\}_j\in\Sigma_m\ \forall i\\ \sigma_1,\ldots,\sigma_n\in\mathcal{P}(\Omega)}} \sum_{i=1}^{n_s} \mathrm{Loss}\left(\rho_i, W_2\mathrm{Bar}(\{\omega_j(\rho_i);\sigma_j\}_{j=1}^m)\right). \tag{5.45}$$

Learning the atoms from data requires the solution of a complicated multilevel optimization problem in Equation (5.45). After entropic regularization, the gradients involved can be computed from the closed formulas available ([117], Section 3.2) or by automatic differentiation techniques.

Note that the gradient of $\rho \mapsto W_2(\rho,\sigma)^2$ is a by-product of the Sinkhorn algorithm in any case. The time needed to solve for weights and atoms can be drastically reduced by employing a warm-start method: when using a quasi-Newton method, instead of running the Sinkhorn algorithm to convergence at every Newton iteration, one can calculate the value of

$$(\{\omega_i;\sigma_i\}_i \mapsto \sum_{i=1}^{n_s} \mathrm{Loss}\left(\rho_i, W_2\mathrm{Bar}(\{\omega_j(\rho_i);\sigma_j\}_{j=1}^n)\right) \tag{5.46}$$

by using only a few Sinkhorn iterations, save the scaling factors $\{\mathtt{a}_{(i)}^{(l)}, \mathtt{b}_{(i)}^{(l)}\}_i$, update the weights and atoms, and then perform another small number of Sinkhorn iterations, starting from the scaling factors $\{\mathtt{a}_{(i)}^{(l)}, \mathtt{b}_{(i)}^{(l)}\}_i$. By interweaving the two iterative procedures in this way, the authors achieve large speed-up at comparable accuracy even when doing as few as two Sinkhorn iterations per Newton step.

Sparsity

Another contribution of [92] is the introduction of a *sparse regularizer*. The authors introduce the modified problem

$$\min_{\substack{\{\omega_j(\rho_i)\}_j\in\Sigma_m\ \forall i\\ \sigma_1,\ldots,\sigma_n\in\mathcal{P}(\Omega)}} \sum_{i=1}^{n_s} W_2\left(\rho_i, W_2\mathrm{Bar}(\{\omega_j(\rho_i);\sigma_j\}_{j=1}^n)\right)^2 + \alpha\sum_{i=1}^{n_s}\sum_{j=1}^{m}\omega_j(\rho_i)W_2(\rho_i,\sigma_j)^2 \tag{5.47}$$

with a hyperparameter $\alpha > 0$. The added term promotes non-zero weights $\omega_j(\rho_i)$ for the nearest neighbor of ρ_i in the set of atoms σ_j. In the limit of $\alpha \to \infty$, for fixed atoms, the optimal weight vector for ρ_i only has one non-zero element for $j^* = \arg\min_j W_2(\rho_i,\sigma_j)^2$. The approach is called *soft Wasserstein K-means* accordingly, in analogy to the K-means clustering method which assigns to each observation the nearest of K *cluster points*.

Sparsity in this sense is not only beneficial for the memory footprint and run-time and of the reconstruction problem $\{\omega_i^\rho;\sigma_i\}_i \mapsto W_2\mathrm{Bar}(\{\omega_i^\rho;\sigma_i\}_i)$, it also promotes uniqueness of the solution to Equation (5.45). Indeed, consider the following example:

Example 5.2 (Non-uniqueness of the Wasserstein dictionary learning problem)**.** *Assume that all input densities $\{\rho_i\}_i$ lie on a subset of the displacement interpolant between two densities ρ_1^* and ρ_2^*, i.e.*

$$\rho_i = \rho(t_i) = ((1-t_i)\mathrm{id}) + t_i T_{\rho_1^*\to\rho_2^*})_\sharp\rho_1^* \tag{5.48}$$

for $t_i \in [\frac{1}{4},\frac{3}{4}]$. In this case, any set of measures $\{\sigma_j\}_{j=1}^m = \{\rho(t_j)\}_{j=1}^m$ lead to the same minimum in Equation (5.45) *as long as at least one element of $\{t_j\}_j$ lies in $[0,\frac{1}{4}]$ and $[\frac{3}{4},1]$, respectively.*

Additional redundant atoms along the interpolant can be added without affecting the optimality. Both of these issues are remedied by the added term in Equation (5.47)*. After the modification, the unique optimal atoms are given by $\sigma_1 = \rho_{t=1/4}$ and $\sigma_2 = \rho_{t=3/4}$. However, how to rigorously extend this result to the case of data that does not lie on a single geodesic through $(\mathcal{P}(\Omega), W_2)$ remains an open problem [92].*

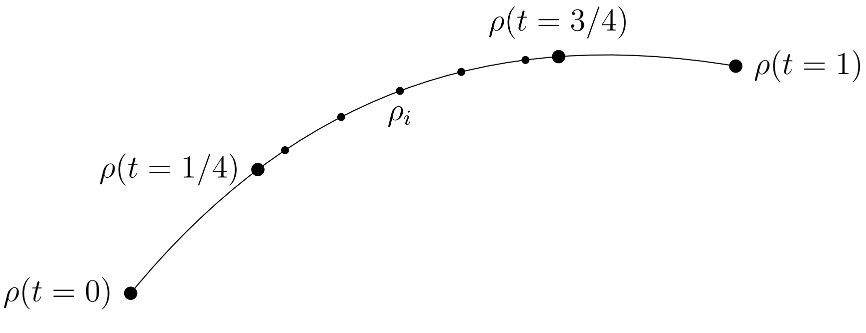

Figure 5.1: Illustration of Example 5.2.

Applications

There is a steadily growing body of literature on the barycentric encoding method; we can not cover the existing approaches in their entirety here. An incomplete list of application cases are computer graphics (color transfer, representation of 3d shapes, reflectance data) [27], images from medical and biological applications (MRI scans, cardiac sequences, facial expressions) [27, 117], correction of telescope images [117], encoding of hand-written digits [117, 92, 143], semantic clustering of texts [117, 144, 124, 42, 80, 92], and building interpolating models for general parametric or time-series data [133, 40].

5.3.7 Parametrized partial differential equations

In this section, we provide a summary of the application of methods based on interpolation and convex combinations in $(\mathcal{P}(\Omega), W_2)$ to PPDE problems.

Convex displacement interpolation

Convex displacement interpolation (CDI), introduced in [73], is based on a linear approximation of the displacement interpolation between two measures. Given one (scalar) parameter $s \in [0,1]$, the CDI provides a mapping

$$s \mapsto \mathrm{CDI}(s, u(0), u(1)) \approx u(s). \tag{5.49}$$

This approximation is motivated by the fact that displacement interpolation between $u(0)$ and $u(1)$ is close - or in some cases identical to - the solution of a PDE problem with initial condition $u(0)$ on the time interval $t \in [0,T]$ with an appropriate rescaling $t \mapsto s(t)$. In order to solve the OT problems needed to build the displacement interpolations, the authors rely on the closed form available for the transport between multivariate Gaussian densities given in Section 3.9.

In cases where the solution to the PDE problem at hand is not itself a probability measure, the authors propose a method to identify coherent features of the solution using a suitable testing function that will be discussed in Section 6.1. Points in the domain where the testing function returns, for example, positive values are interpreted as independent identically distributed realizations of a multivariate Gaussian distribution.

Greedy barycenter methods

In [56], barycentric encoding is used to obtain reduced models for PPDE problems where the solutions can be represented by probability densities, namely the Burger's equation, Camassa-Holm equation, and Korteveg-de-Vries equation.

In particular, given a set of atoms $\{\sigma_j\}_j = \{\rho(\mu_j)\}_j$ where $\mathcal{L}(\mu_j, \rho(\mu_j)) = 0$ for some parameter value $\mu_j \in \mathcal{A}$, the solution to the parametrized PDE problem for $\mu^* \in \mathcal{A}$ is expressed as

$$\rho(\mu^*) \approx W_2\mathrm{Bar}(\{\omega_j(\mu^*); \sigma_j\}_{j=1}^m). \tag{5.50}$$

Just as in a number of the applications in the previous section, the optimal transport barycenter is used as a non-linear interpolation method between snapshots of PPDE solutions.

Note that finding the optimal weights $\omega_j(\mu^*)$ by solving the encoding problem (5.25) requires knowledge of $\rho(\mu^*)$, which is not available in the online phase. Therefore, the mapping $\mu \mapsto \omega_j(\mu)\ \forall j = 1, \dots, m$ is approximated by interpolation.

In the offline phase, the PPDE is solved for the training values $\{\mu_i\}_{i=1}^{n_s} \subset \mathcal{A}$. Then, the optimal weights $\omega_j(\mu_i)\ \forall j = 1, \dots, m$ and $i = 1, \dots, n_s$ are obtained by solving the barycenter encoding problem (5.42). The pairs $\{\mu_i, \{\omega_j(\mu_i)\}_j\}_i$ are then used to build the interpolation $\mu \mapsto \{\omega_j\}_j$.

The online phase consists of evaluating this interpolating function and computing the corresponding barycenter, which approximates $\rho(\mu^*)$.

In order to obtain the atoms, a greedy algorithm as in Section 2.3 is used. For every element of the training set $\{\rho(\mu_i)\}_{i=1}^{n_s}$, the best approximation using optimal barycentric coordinates given the present set of atoms is calculated. Then, the element of the training set with the largest approximation error is added as an additional atom until a given tolerance is reached.

Assuming m snapshots $\{\rho(\mu_{i^1}), \dots, \rho(\mu_{i^m})\} =: \{\sigma_j\}_{j=1}^m$ have been selected, one lets

$$i^{m+1} = \underset{1 \le i \le n_s}{\arg\max} \min_{\{\omega_j\}_{j=1}^m \in \Sigma_m} \text{Loss}\left(\rho(\mu_i), W_2\text{Bar}(\{\omega_j; \sigma_j\}_{j=1}^m)\right) \tag{5.51}$$

and $\rho(\mu_{i^{m+1}}) =: \sigma_{m+1}$ is added to the set of atoms.

Since the examples in [56, 14] take place in one dimension, the calculations are greatly simplified. Recall that in one spatial dimension, there is a closed formula available for the transport from ρ to σ, given by their cumulative distribution functions (cdf):

$$T_{\rho\to\sigma} = \text{cdf}(\sigma)^{[-1]} \circ \text{cdf}(\rho).$$

When $\Omega = [0,1]$ and $\bar{\rho} = \mathbb{1}_{[0,1]}$, the Monge embedding coincides with the inverse cdf operation, i.e. $T_{\bar{\rho}\to\sigma} = \text{cdf}(\sigma)^{[-1]}$. For any $\rho, \sigma \in \mathcal{P}(\mathbb{R})$

$$W_2(\rho, \sigma) = \|\text{cdf}(\rho)^{[-1]} - \text{cdf}(\sigma)^{[-1]}\|_{L^2([0,1])}, \tag{5.52}$$

and the optimal transport barycenter has the closed form

$$(\text{cdf} \circ W_2\text{Bar}(\{\omega_i; \sigma_i\}_{i=1}^m))^{[-1]} = \sum_{i=1}^m \omega_i \,\text{cdf}(\sigma_i)^{[-1]}, \tag{5.53}$$

which is a special case of Definition 5.3. If one lets Loss $= W_2^2$ in Equation (5.25), the barycenter encoding problem is therefore a quadratic optimization problem over the convex set Σ_m:

$$\{\omega_i^\rho\}_i = \underset{\{\omega_i\}_i \in \Sigma_m}{\arg\min} \left\| \text{cdf}(\rho)^{[-1]} - \sum_{i=1}^m \omega_i \,\text{cdf}(\sigma_i)^{[-1]} \right\|_{L^2([0,1])}^2 . \tag{5.54}$$

Sparse barycenters and local Euclidean embedding

The approach from [56, 14] is extended in [51] in two ways. Firstly, the method is used to treat a two-dimensional Burger's equation by using the debiased entropic optimal transport barycenters from Definition 5.2. The barycenter encoding and reconstruction are solved using the iterative methods introduced in Section 4.3 with automatic differentiation techniques to obtain the needed gradients.

Secondly, instead of relying on one fixed set of atoms $\{\sigma_i\}_{i=1}^m$, the method introduces the notion of *best m-term barycenter.* Here, the set of atoms is the entire training set $\{\rho_i\}_{i=1}^{n_s}$, but for every encoding, only m of the entries of $\{\omega_i\}_{i=1}^{n_s}$ are nonzero. The set of such m-sparse weight vectors is denoted $\Sigma_{n_s}^m$, which is a collection of $\binom{n_s}{m}$ m-simplices.

While the approximation error with an m-sparse barycenter is necessarily equal or worse than one using the entire training set as atoms (since the former constitutes a subset of the latter), it is preferable in practice. The sparse approach saves computational time in the reconstruction step, which linearly scales with the number of atoms. Furthermore, too large a set of atoms might introduce redundancies and local minima in the encoding problem. Lastly, approximation by an optimal m-sparse barycenter will always be better or equal than a barycenter using m atoms constructed by a greedy method.

Since the problem of finding an optimal m-sparse barycenter corresponding to a parameter value $\mu^* \in \mathcal{A}$ as in Equation (5.50) requires knowledge of $\rho(\mu^*)$, a surrogate needs to be built for the mapping $\mu \mapsto \{\omega_j^{\rho(\mu)}\}_j \in \Sigma_{n_s}^m$ from parameters to optimal sparse barycenter weights. This is done by learning a function

$$\text{approxdist} : \mathcal{A} \times \mathcal{A} \to \mathbb{R} : \text{approxdist}(\mu_1, \mu_2) \approx W_2(\rho(\mu_1), \rho(\mu_2))^2. \tag{5.55}$$

Then, the sparse weights are optimized until

$$W_2(\rho(\mu_i), W_2\text{Bar}(\{\omega_j(\mu^*); \rho_j\}_j))^2 \approx W_2(\rho(\mu_i), \rho(\mu^*))^2 \approx \text{approxdist}(\mu_i, \mu^*) \tag{5.56}$$

for all parameter values μ_i in the training set.

This construction is motivated by the fact that in Euclidean spaces, given a point in an m-simplex (corresponding to $\rho(\mu^*)$), the position of this point (given by $\{\omega_i(\mu^*)\}_i$) is uniquely determined by its distance to all vertices of the simplex (corresponding to $\{\rho(\mu_i)\}_i$).

5.4 Application to a porous media equation

In this section, we recall some results from [14], where we applied the methodology from [56] that was described in Section 5.3.7 to a one-dimensional porous media equation as it is used in hydrology and reservoir engineering. We will use the notation from [14], which might lead to some duplications of symbols that were used in other parts of this thesis, but their meaning should be clear from the context.

Described physical system

The system describes the flow of the *wetting saturation* s_w, which follows a non-linear continuity equation of the form

$$\phi(x)\partial_t s_{\mathrm{w}}(x,t) + \partial_x \left(\frac{\lambda_{\mathrm{w}}(s_{\mathrm{w}}(x,t))\, v(x,t)}{\lambda_{\mathrm{w}}(s_{\mathrm{w}}(x,t)) + \lambda_{\mathrm{nw}}(1 - s_{\mathrm{w}}(x,t))} \right) = 0 \quad \forall x \in [0,1], t \in [0, t_F], \tag{5.57}$$

where $\phi \in (0,1]$ is the porosity of the medium, The total Darcy velocity v is given by

$$v(x,t) = (\lambda_{\mathrm{w}} + \lambda_{\mathrm{nw}}) k(x) \partial_x p(x,t) \tag{5.58}$$

and subject to the constraint $\partial_x v(x,t) = 0$, i.e. it is defined via an elliptic problem for the pressure p. The permeability of the medium is denoted by k, while λ_{w} and λ_{nw} are the phase mobilities of s_{w} and $s_{\mathrm{nw}} = 1 - s_{\mathrm{w}}$, which depend on the phase viscosities μ_{w} and μ_{nw} through the power law

$$\lambda_{\mathrm{w}} = \mu_{\mathrm{w}}^{-1} s_{\mathrm{w}}^{\beta} \text{ and } \lambda_{\mathrm{nw}} = \mu_{\mathrm{nw}}^{-1}(1 - s_{\mathrm{w}})^{\beta} \tag{5.59}$$

with $\beta \in \mathbb{R}_{>0}$.

The system is accompanied by constant Dirichlet boundary conditions for the pressure (with $p(0,t) > p(L,t)$ to drive the flow) and the datum $s_{\mathrm{w}}(0,t) = 1 \quad \forall t \in [0, t_F]$. The initial condition is $s_{\mathrm{w}}(x,0)|_{x>0} = 0$.

For the example we will present here, ϕ, k, and μ_{nw} are kept constant while $\beta, \mu := \mu_{\mathrm{nw}}/\mu_{\mathrm{w}}$, and t are treated as varying parameters.

Offline and online phase

The high fidelity solutions are computed using a finite-volume method with implicit time-stepping for the pressure equation and explicit time-stepping for the saturation evolution equation. To respect the CFL condition, the time-step size in the latter is adaptive. The spatial discretization consists of $N = 1002$ cells.

The solutions s_{w} are not probability densities since the Dirichlet boundary condition acts like a source term. Therefore, to apply the procedure outline above, the snapshots s_{w} are first normalized to obtain densities

$$\rho(\beta,\mu,t) := \frac{s_{\mathrm{w}}(\beta,\mu,t)}{m_{\mathrm{w}}(\beta,\mu,t)} := \frac{s_{\mathrm{w}}(\beta,\mu,t)}{\int_{[0,1]} s_{\mathrm{w}}(\beta,\mu,t)}. \tag{5.60}$$

As for the barycenter weights, an interpolation is built for the map $(\beta,\mu,t) \mapsto m(\beta,\mu,t) := \int_{[0,1]} s_{\mathrm{w}}(\beta,\mu,t)$.

In the online phase, given μ^*, we approximate $\rho(\mu^*)$ using a barycenter approximation and the interpolation $(\beta,\mu,t) \mapsto \{\omega_i(\mu^*)\}_i$. In particular, a linear interpolation and subsequent projection to Σ_m is used. The mass is recovered by evaluating the interpolation function $(\beta,\mu,t) \mapsto m_{\mathrm{w}}$.

Example problem

In the example we present here, the training set consists of the densities

$$\begin{aligned} &\{\rho(\beta,\mu,t)_i\}_{i=1}^{n_s} \subset L^\infty([0,1]) : \\ &\qquad \mu \in \{1,2,3,6,12,25\}, \beta \in \{2,3,4,5,6\}, t \in \{0.2, 0.4, \dots, 5\}, \end{aligned} \tag{5.61}$$

hence $n_s = 750$. As shown in Figure 5.2, the solutions are characterized by a moving front of s_{w}, whose shape and speed changes with β and μ.

Greedy dictionary construction

The results of the greedy algorithm are shown in Figure 5.3 and Figure 5.4. The algorithm is initialized with those two elements of $\{\rho(\beta,\mu,t)_i\}_{i=1}^{n_s}$ as atoms that have the largest distance to one another, measured in the W_2 norm.

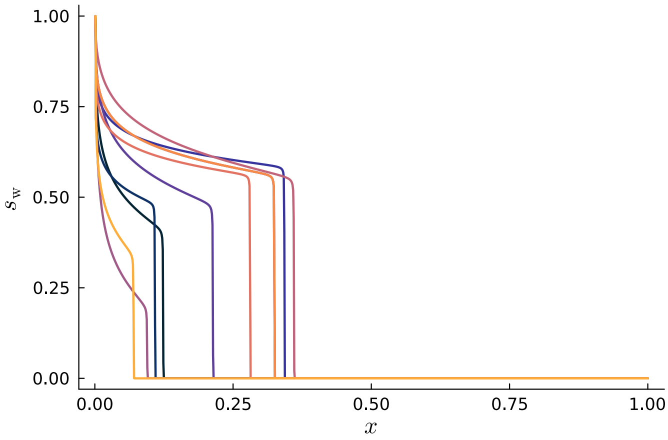

Figure 5.2: Elements of the training set $\{\rho(\beta,\mu,t)_i\}_{i=1}^{n_s}$.

If the minimization algorithm fails to converge for a specific snapshot ρ_i, then the approximation from a previous iteration is used and ρ_i is added to the set of atoms. This does happen in practice, since the greedy algorithm is only a heuristic to avoid redundancies in the set of atoms. This issue is discussed in more detail below.

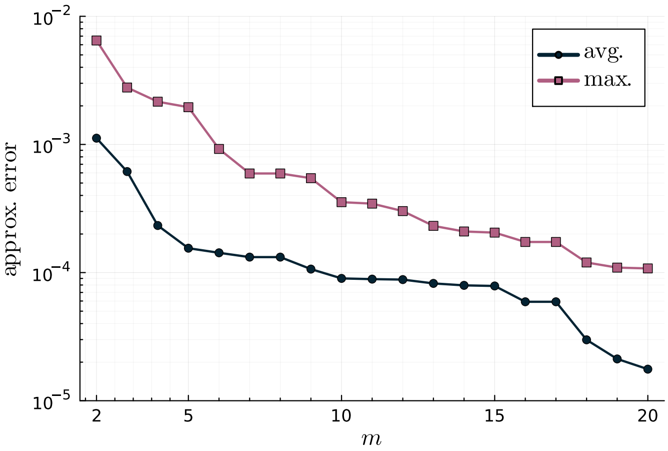

Figure 5.3: Evolution of the average and maximum of $\{\min_{\{\omega_j\}_{j=1}^m \in \Sigma_m} \text{Loss}\left(\rho(\mu_i), W_2\text{Bar}(\{\omega_j;\sigma_j\}_{j=1}^m)\right)\}_{i=1}^{n_s}$ throughout the iterations of the greedy algorithm where Loss $= W_2$.

Reconstruction errors

In Figure 5.5, we show two reconstructions of s_w for $m = 5$ and compare them to their projection to a POD basis with $n = 50$. The POD approximation exhibits the expected shortcomings when approximating functions with a jump discontinuity. In contrast, the approximation based on the optimal transport barycenter exhibits the characteristic wavefront without spurious oscillations. In the worst-case approximation, there is a notable discrepancy close to $x = 0$, where the barycenter approximation greatly underestimates the value of the saturation. This is an artifact of the normalizing and a consequence of the lacking resolution of the mass

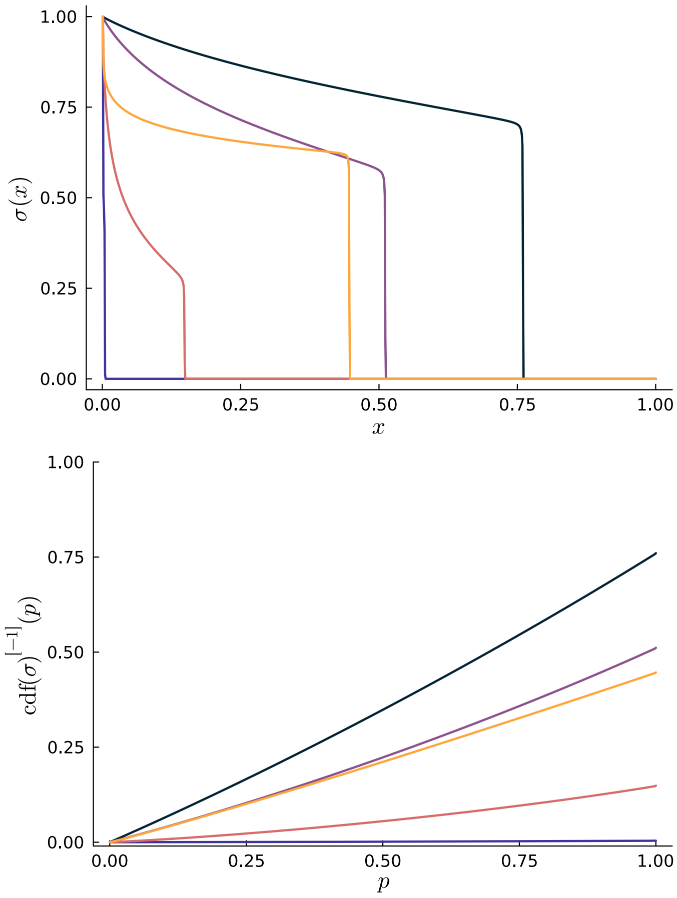

Figure 5.4: The first five atoms selected by the greedy algorithm (top, from dark to bright) and their corresponding inverse cumulative distribution functions (bottom).

interpolation function for solutions with very small mass. One could circumvent this problem by enforcing the value $s_{\mathrm{w}}|_{x=0} = 1$ in the reconstruction step instead of relying on an interpolation. However, the latter approach is more general.

Properties of the minimization problem

Next, we consider the energy landscape of the minimization problem in Figure 5.6. The values of the W_2 distance between the density corresponding to the saturation on the left in Figure 5.5 and its approximation by an optimal transport barycenter of $m = 3$ atoms is plotted for different weight values, corresponding to the barycentric coordinates on the pictured triangle.

We see that the contour lines of the loss function are very eccentric and that there are several weight vectors that give nearly identical reconstruction errors. This is confirmed when we consider the function

$$(\delta\omega_2, \delta\omega_3) \mapsto W_2(\rho(\mu^*), W_2\mathrm{Bar}(\omega_1(\mu^*) - \delta\omega_2 - \delta\omega_3, \omega_2(\mu^*) + \delta\omega_2, \omega_3(\mu^*) + \delta\omega_3; \sigma_1, \sigma_2, \sigma_3)) \quad (5.62)$$

that describes the approximation error in the neighborhood of the optimal weights $\{\omega_i(\mu^*)\}_{i=1}^m$. This energy landscape is depicted in Figure 5.6 in $\log_{10}$ scale. From

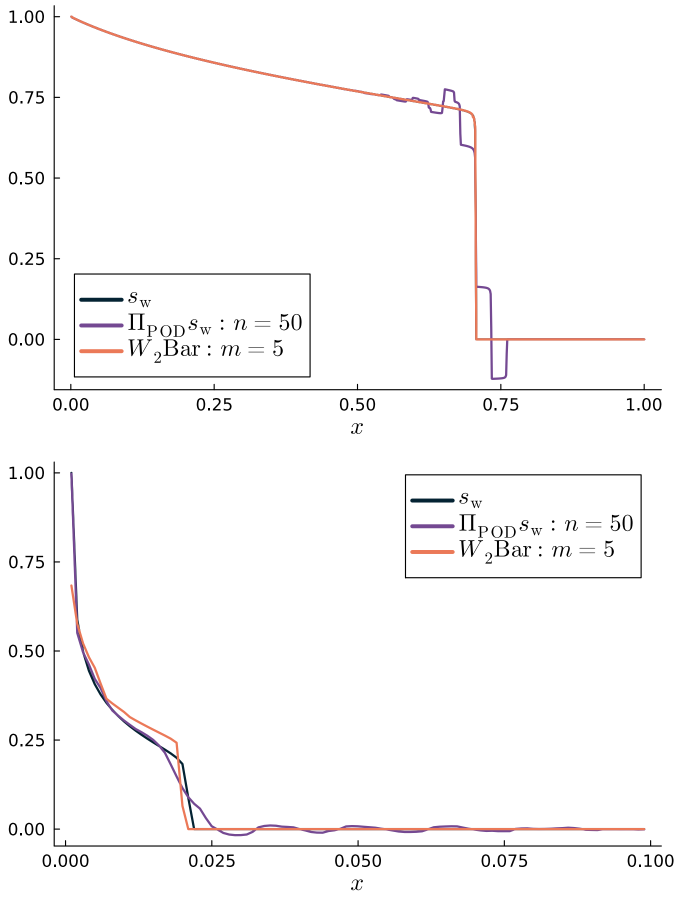

Figure 5.5: The best (top) and worst (bottom) approximation from the training set for $t > 1$, compared with a POD projection. Note that the lower figure only shows a fraction of the domain.

these results we can conclude that the greedy algorithm is susceptible to creating redundancy in the set of atoms.

Atom redundancy

Lastly, we use the fact that in one spatial dimension, the set of possible barycenters obtained from atoms $\{\sigma_i\}_i$ (c.f. Equation (5.25)) is given by the convex hull of $\{\mathrm{cdf}(\sigma_i)^{[-1]}\}_i$, who are elements of a Hilbert space and form a simplex. The volume of this simplex can be computed using the Cayley-Menger determinant. Figure 5.7 shows how this volume, normalized with the volume of the unit m-simplex $(m!)^{-1}$, changes as the greedy algorithm adds atoms to the dictionary.

We see that the volume decreases exponentially at a fast rate. Note that the following recursive relation holds for the volume of two subsequent normalized simplices:

$$\frac{\mathrm{Vol}(\Sigma(\{\mathrm{cdf}(\sigma_i)^{[-1]}\}_{i=1}^{m+1}))}{\mathrm{Vol}(\Sigma_{m+1}}) = \delta^{m+1}\frac{\mathrm{Vol}(\Sigma(\{\mathrm{cdf}(\sigma_i)^{[-1]}\}_{i=1}^{m}))}{\mathrm{Vol}(\Sigma_m}, \tag{5.63}$$

where δ^{m+1} is the orthogonal distance of $\mathrm{cdf}(\sigma_{m+1})^{[-1]}$ to the simplex after m iterations. We can conclude that many of the atoms added in the later greedy

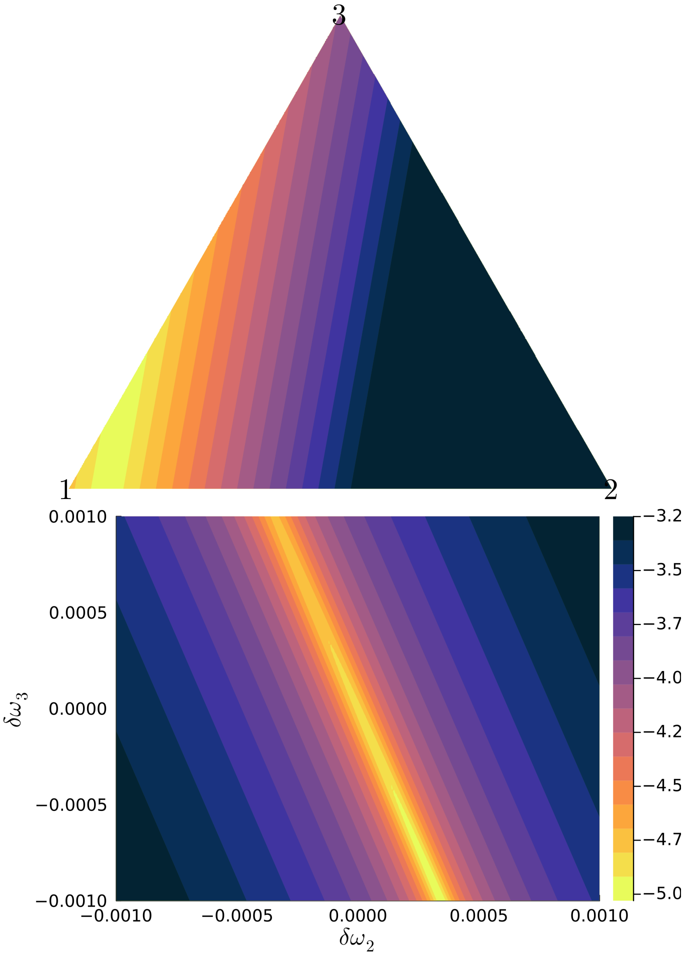

Figure 5.6: Values of the approximation error to the density corresponding to the saturation on the left in Figure 5.5. Top: Values of the error for all possible weight vectors. The colorbar is linear in $[0, 0.2]$. Bottom: Values around the optimal weight vector $\omega_{1,2,3}(\mu^*) \approx (0.83, 0.03, 0.13)$ in $\log_{10}$ scale.

iterations lie very close to the simplex generated in the previous iterations, i.e. δ^{m+1} is very small.

Further details on the implementation of the optimization algorithm, the interpolation methods, and further numerical examples can be found in the original reference [14].

5.5 Conclusion

Methods inspired by and using results from optimal transportation theory to approximate data in the space of probability measures have been developed in great number in recent years. This has been facilitated by the development of fast methods for computational optimal transport like the ones we described in Chapter 4.

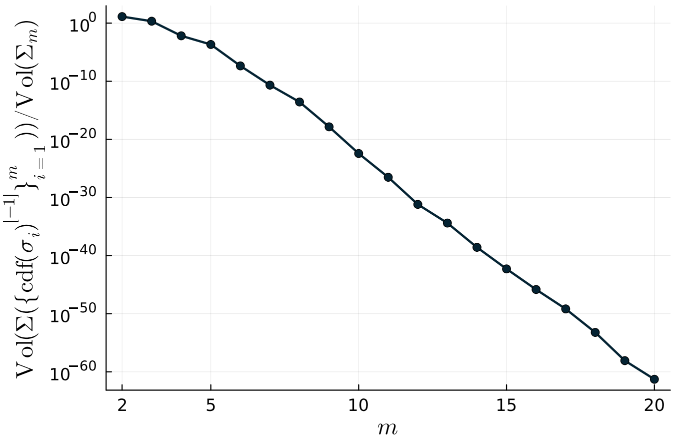

Figure 5.7: Normalized volume of the convex hull of the dictionary atoms through the iterations of the greedy algorithm.

Interpolation and residual-based methods

Concerning the application to the model order reduction of parametrized partial differential equations, the methods in [73, 56, 51, 14, 133] all rely on interpolation methods in their online phase. If one wants to avoid this, the optimal weights can be obtained by substituting Equation (5.50) directly into the PPDE problem to solve, for example,

$$\min_{\{\omega_j\}_j \in \Sigma_m} \left| \mathcal{L}_\mu \left(W_2\mathrm{Bar}(\{\omega_j; \sigma_j\}_{j=1}^m) \right) \right|^2 . \tag{5.64}$$

The difficulty here lies in the cost, which is restrictive for the use in a reduced model online phase, since the barycenter has to be reconstructed from the weights and the PDE residual has to be evaluated in the high-fidelity space.

Encoding uniqueness

Another open question is how to handle the issue of redundancy in the set of atoms. Both the greedy and the POD algorithm from the reduced basis method rely on orthogonalization in order to keep the projection onto the reduced basis well-conditioned. While the greedy barycenter algorithm from [56, 14] provides a heuristic to avoid elements to the set of atoms that are already within reach of the dictionary, we see in practice that the optimal weights are not unique (c.f. Figure 5.6). The adaptive method from [51] shares this difficulty, see in particular Figure 6(b) therein.

Wasserstein dictionary learning approaches [117, 92] and in particular the additional term to enforce sparsity in the weights in [92] might be able to overcome these issues. However, determining the atoms by optimization leads to a highly non-linear optimization problem that is both multi-level and non-convex. When optimizing for both the weights and atoms simultaneously, it is possible to reach local minima where, for example, the weights are optimal for the current iteration of atoms. In this case, it is necessary to restart the optimization and to introduce additional weighting hyperparameters in order to force the optimization to focus on either atoms or weights. While such tricky optimization problems are not uncommon in dictionary learning or machine learning in general, they still constitute a challenge to implementation of these methods.

Chapter 6

Transport-based registration

We have seen in Chapter 5 that the optimal transport barycenter approximation achieves good accuracy with already very few atoms. However, when trying to push this accuracy further, one has to face issues of redundancy in the set of atoms and the rising cost of the reconstruction problem.

In this section, we will present the method from [25], which uses the ideas from the barycenter encoding to build a more classical registration method as discussed in Section 2.6. As a result, the content of this section and the cited work by the author are largely identical.

Given a set of densities $\{\rho(\mu_i) = \rho_i\}_{i=1}^{n_s}$, we fix a reference density $\bar{\rho}$ (this can, but does not have to be, a weighted barycenter of the ρ_is). Then, the maps $T_{\bar{\rho}\to\rho_i} = \mathrm{ME}_{\bar{\rho}}(\rho_i)$ are computed for all i.

The idea is now to use these maps as a starting point for the construction of the registration map Φ_μ that should inherit the good interpolation properties of the displacement interpolation and barycenter encoding. In order to resolve the solution further, we rely on a classical reduced basis in the reference frame, i.e. after the registration step.

Remark 6.1. *In this section, all optimal transport quantities ($\psi, \psi^c, T, \dots$) are denoted by their un-regularized form. The proposed method is applicable in this setting under sufficient regularity assumptions discussed in Section 3.4.*

For computational feasibility, we use entropic regularization in the numerical examples of Section 6.8. To apply the method in this case, one has to make the appropriate replacements, i.e. the c-transform becomes an application of the softmin, W_2 becomes $W_{2,\varepsilon}$ or S_ε, $\psi_{\rho\to\sigma}$ becomes either $\psi^\varepsilon_{\rho\to\sigma}$ or $\psi^\varepsilon_{\rho\to\sigma} - \psi^\varepsilon_{\rho\to\rho}$, and so forth. When specific assumptions or steps change depending on what regularisation is used, it is explicitly stated.

6.1 Dimension reduction on the Monge embeddings

We begin our presentation with the strategy to obtain a small number of *transport modes* that are the basis for registration. Consider the following two motivating examples.

Example 6.1. *Recall the case $\mathcal{M} = \{x \mapsto u_0(x-\mu) : \mu \in \mathbb{R}\}$ from Example 2.4, the solution manifold of a pure equation with slow $n^{-1/2}$ n-width decay. If $u_0 \in L^1(\mathbb{R})$ is normalized (and this can be done without loss of generality), we can choose $\bar{\rho} = u_0$ and let $\rho(\mu) := u_0(x-\mu)$. The Monge embeddings of the set of solutions are of the form*

$$\mathrm{ME}_{\bar{\rho}}(\rho(\mu)) = \mathrm{id} + \mu \tag{6.1}$$

with the corresponding potentials $\{\psi^c_\mu : y \mapsto -\mu y\}$.

Example 6.2. *Let $s \in L^1(\mathbb{R}_{\geq 0})$ be a non-increasing function with compact support such that $\int_0^{+\infty} s(x)\mathrm{d}x = 1$. Consider the set $\mathcal{M} = \{x \mapsto s(x/\mu) : \mu \in \mathbb{R}_{>0}\}$. Let $\bar{\rho} = s$ and $\rho(\mu) := \mu^{-1}s(x/\mu)$. The Monge embeddings of elements of $\mathcal{M}$ take the form*

$$\mathrm{ME}_{\bar{\rho}}(\rho(\mu)) = \mu\,\mathrm{id} \tag{6.2}$$

with the corresponding potentials $\{\psi^c_\mu : y \mapsto \frac{1}{2}\mu y^2\}$.

Remark 6.2. *In Example 6.2, we can take $s(y) = \mathbb{1}_{[0,1]}(y)$ or $s(y) = C\exp(-1/(1-y^2))\mathbb{1}_{[0,1]}$ with normalizing factor C. Note that these functions are similar in shape to the solutions of the porous media equation from Section 5.4. As such, they are very hard to approximate using a linear reduction method.*

These examples illustrate that the set of Monge embeddings is extremely easy to approximate for the cases of translations and dilations, while the corresponding solutions themselves are not.

Consider a PPDE problem with solution $u(\mu)$ and related densities $\rho(u)(\mu) =: \rho(\mu)$. In some cases, $\rho(u) = u$ is a possible choice, as in Example 6.1. The requirements for ρ are that it returns probability densities that coincide with the features of the solution that have to be registered.

Example 6.3. *In [73], a scalar testing function $\mathcal{T}$ is chosen to determine the distribution of features (which are sets of points $\{x \in \Omega : \mathcal{T}(x;u) > 0\}$). Examples for $\mathcal{T}$ considered therein include $\|\nabla \times u\|, \|\nabla u\|$, the derivative of the Mach number of a flow, and a shock discontinuity indicator (Equation (34) therein).*

For the proposed method, ρ should return continuous densities supported on the entire domain to allow the application of the regularity results presented in Section 3.4.

Remark 6.3. *The discretization of u and ρ need not agree. It can be beneficial to discretise the latter on a regular tensor grid to accelerate the computation of the transport mappings, c.f. Section 4.2.*

Transport mode construction

We now describe the proposed method, assuming that we have access to a set of snapshots $\{u(\mu_i)\}_{i=1}^{n_s} \subset V_h$ that are solutions to some high-fidelity discretization of the PPDE in question at different parameter values.

First, compute $\{\rho(u)(\mu_i)\}_{i=1}^{n_s}$ and denote by $\bar{\rho}$ a suited reference density, e.g. $\rho(\bar{\mu})$ for a certain parameter value $\bar{\mu}$, or a weighted optimal transport barycenter of $\{\rho(\mu_i)\}_{i=1}^{n_s}$. Next, calculate the Monge embeddings $\{T_{\bar{\rho}\to\rho(\mu_i)}\}_{i=1}^{n_s}$. We denote by ψ^c_i the transport potential such that $T_{\bar{\rho}\to\rho(\mu_i)}(y) = y - \nabla\psi^c_i(y)$.

Definition 6.1. *The* transport modes *of a set of probability measures $\{\rho(\mu_i)\}_{i=1}^{n_s}$ and a reference $\bar{\rho}$ are given by*

$$y \mapsto \xi_j^c(y) = (\lambda_j^\psi)^{-1/2} \sum_{i=1}^{n_s} (\mathrm{v}_j^\psi)_i \psi_i^c(y), \tag{6.3}$$

where λ_j^ψ and v_j^ψ are jth non-zero eigenvalue and eigenvector of the Monge embedding correlation matrix

$$\mathbb{C}^\psi := \{\langle \nabla\psi_i^c, \nabla\psi_j^c\rangle_{L^2(\bar{\rho})}\}_{1\le i,j\le n_s}. \tag{6.4}$$

Note the similarities with Algorithm 1 for the POD modes.

Example 6.4. *For snapshots of the pure advection equation from Example 6.1 at different values μ_i, $\{\bar{\rho}(x-\mu_i)\}_{i=1}^{n_s}$, we find $(\mathbb{C}^\psi)_{ij} = \mu_i\mu_j$ with one non-zero eigenvalue $\lambda_1^\psi = \sum_{i=1}^{n_s} \mu_i^2$ and eigenvector $(\mathrm{v}_1^\psi)_i = \mu_i(\sum_{j=1}^{n_s} \mu_j^2)^{-1/2}$. The corresponding transport mode is given by $\xi_1^c : y \mapsto -y$.*

The case of the moving front in Example 6.2 is almost identical, as $(\mathbb{C}^\psi)_{ij} = \mu_i\mu_j \int y^2 \mathrm{d}\bar{\rho}(y)$. Also in this case, there is only one transport mode which is proportional to $y \mapsto y^2$.

If the eigenvalues of $\mathbb{C}^\psi$ decay fast enough, all transport potentials $\psi^c(\mu)$ can be accurately (in the sense of a $\bar{\rho}$-weighted L^2 norm of their derivatives) approximated by a linear combination of the form $\psi^c(\mu) \approx \sum_{j=1}^m w_j(\mu)\xi_j^c$ where $m \ll n_s$.

Limits of the linear treatment

Note that we cannot completely escape the non-linearity of $\mathcal{P}(\Omega)$. In order to guarantee that the approximate transport potential $\sum_{j=1}^m w_j(\mu)\xi_j^c$ is in fact a transport potential (i.e.: convex) it would be sufficient to take convex combinations of the snapshot potentials: $\psi^c(\mu) \approx \sum_{i=1}^{n_s} \omega_i(\mu)\psi_i^c$ with non-negative weights $\omega_{1,\dots,m}$ that sum to one. In that case, the function $\frac{1}{2}|y|^2 - \sum_{i=1}^{n_s} \omega_i(\mu)\psi_i^c$ is again a convex function, the same one that appears in the definition of the Monge embeddings barycenter.

Using a linear combination of transport modes, we expect the resulting function $\frac{1}{2}|y|^2 - \sum_{j=1}^m w_j(\mu)\xi_j^c(y)$ still to be convex, but this is not guaranteed by construction but a consequence of the quality of approximation through the modes $\{\xi_i^c\}_{i=1}^m$.

6.2 Reference reduced basis

Evaluating

$$u(\mu_i) \circ \left(\mathrm{id} - \nabla \sum_{j=1}^m w_j(\mu_i)\xi_j^c\right) =: u(\mu_i) \circ \Phi_{\mu_i}^{-1} \tag{6.5}$$

applies the approximated transport mapping to the ith snapshot and yields elements of the mapped snapshot manifold $\Phi_\mu(\mathcal{M})$. By construction, we expect this set to be more amicable to linear approximation. Returning to the simple cases of Example 6.1 and Example 6.1, there is only one transport mode each. They have the form $\xi_1^c(y) = -y$ and $\xi_1^c(y) = \text{const.}\, y^2$. The approximation of the snapshot potentials $\psi_i^c = -\mu_i y \in \operatorname{span}\{\xi_1^c\}\ \forall i$ and $\psi_i^c = \frac{1}{2}\mu_i y^2 \in \operatorname{span}\{\xi_1^c\}\ \forall i$ is exact in both cases and the mapped snapshot manifold consists of one single element $\bar{\rho}$.

More generally, we proceed by building a reduced basis in the reference space using the correlation matrix of transported snapshots

$$\mathbb{C}^{u\circ\Phi^{-1}} := \{\langle u(\mu_i)\circ\Phi_{\mu_i}^{-1}, u(\mu_j)\circ\Phi_{\mu_j}^{-1}\rangle_{V_h}\}_{1\le i,j\le n_s}. \tag{6.6}$$

Just as in the classical RB method described in Chapter 2, we obtain a set of reduced basis functions which we will denote by $\phi_{1,\dots,n_m}$. Now, any element of $\mathcal{M}$ can be approximated via

$$u_{\mathrm{trb}}(\mu) := \sum_{i=1}^{n_m} \tilde{u}(\mu)_i\,\phi_i \circ \left(\mathrm{id} - \nabla\left[\sum_{j=1}^{m} w_j(\mu)\xi_j^c\right]^c\right) = \sum_{i=1}^{n_m} \tilde{u}(\mu)_i\,\phi_i \circ \Phi_\mu. \tag{6.7}$$

Remark 6.4. *The relation*

$$\Phi_\mu(x) = x - \nabla\left[\sum_{j=1}^{m} w_j(\mu)\xi_j^c\right]^c(x) \tag{6.8}$$

holds as long as $\sum_{j=1}^m w_j(\mu)\xi_j^c$ is in fact a transport potential, i.e. a convex function. In this case, we can use the properties of the c-transform to invert the mapping. This trick is possible since the gradients of Legendre transforms are inverses of each other and the c-transform and Legendre transform are related through Proposition 3.2.

When working with regularized potentials, we still expect that

$$\mathrm{id} - \nabla\psi^{c,\varepsilon} \approx (\mathrm{id} - \nabla\psi^{\varepsilon})^{-1} \tag{6.9}$$

for small ε because of the convergence of the entropic transport map to the transport map of the un-regularized problem as $\varepsilon \to 0$ and the estimates given in Section 4.1.

We conclude this subsection with an example for a one-dimensional PPDE that forms boundary layers from [128]:

Proposition 6.1. *The solutions to the equation*

$$-\partial_{xx}^2 u_\mu + \mu^2 u_\mu = 0 \tag{6.10}$$

on the domain $\Omega = [0,1]$ with boundary conditions $u_\mu(0) = 1$, $u_\mu(1) = 0$ and $\mu, \bar{\mu} \in [\mu_{\min}, \mu_{\max}] =: \mathcal{A}$, $\mu_{\max} = \epsilon^{-2}\mu_{\min}$, $\mu_{\min} > 1$, $\epsilon \in (0,1)$ satisfy

$$\inf_{\xi_1^c\in\mathrm{span}\{\psi_\mu^c:\mu\in\mathcal{A}\}} \sup_{\mu\in\mathcal{A}} \inf_{\substack{w_1(\mu)\in\mathbb{R}\\ \Phi^{-1}(y)=y-w_1(\mu)\partial_y\xi_1^c(y)\\ \Phi_\mu^{-1}:\Omega\to\Omega \text{ is a bijection}}} \|u_{\bar{\mu}} - u_\mu\circ\Phi_\mu^{-1}\|_{L^2(\Omega)} \le e^{-\mu_{\min}}(4+\epsilon). \tag{6.11}$$

where $\rho(u) = u/\int u$, $\bar{\rho} = W_2\mathrm{Bar}(\{\rho_\mu : \mu \in \mathcal{A}\})$, ψ_μ^c denotes the function such that $T_{\bar{\rho}\to\rho_\mu}(y) = y - \partial_y\psi_\mu^c(y)$, and $\int \psi_\mu^c = 0$.

In other words, we can show a bound on the Kolmogorov n-m-width (in the limit of tolerance → 0, c.f. [128], section 3.2) of $\mathcal{M}$ for $n = m = 1$. The proof of this proposition can be found in Appendix C. It relies on the fact that, just as in Example 6.2, the transport mode is very close to the mapping $y \mapsto \bar{\mu}y/\mu$ for $y \le \min\{\mu/\bar{\mu}, 1\}$, and therefore aligns the boundary layers, since $u(\mu,x) \approx exp(-\mu x)$.

6.3 Online phase

The proposed approximation of $u(\mu)$,

$$u_{\mathrm{trb}}(\mu) = \sum_{i=1}^{n_m} \tilde{u}_i(\mu)\, \phi_i \circ \left(\mathrm{id} - \nabla \left[\sum_{j=1}^{m} w_j(\mu) \xi_j^c \right]^c \right)$$

is determined by the values of $\tilde{u}(\mu)_i$ and $w(\mu)_j$ for all $i = 1, \dots, n_m$ and $j = 1, \dots, m$.

Optimization of the basis and mapping coefficients

One option to determine the optimal values of both is to minimize the norm of the PPDE residual with respect to both of these variables. While the dependence of u_{trb} on $\{w_j\}_j$ is non-linear, the Jacobian of this relation has a simple form in the reference frame, see Remark 6.6. The resulting optimization problem in a least squares approach would be of the form

$$\min_{\{\tilde{u}_i\}_i, \{w_j\}_j} \left| \mathcal{L}_\mu \left(\sum_i \tilde{u}_i\, \phi_i \circ \left(\mathrm{id} - \nabla \left[\sum_{j=1}^{m} w_j \xi_j^c \right]^c \right) \right) \right|^2 . \tag{6.12}$$

In this work, we opt for a different approach. The values of the transport mode coefficients is predicted based on the value of μ. Then, the coefficients of the reference frame reduced basis are determined by Galerkin projection.

We learn the mapping $\mu \mapsto w_{1,\dots,m}(\mu)$ using a Gaussian process [110] and the data from the snapshot set $\{\mu_i, w_{1,\dots,m}(\mu_i)\}_{i=1}^{n_s}$. The functions could also be described by interpolation or any related method. We use the Gaussian process as it is computationally cheap also for high-dimensional data.

Online residual assembly

To solve the PPDE problem for a new parameter value μ, we evaluate the mapping Φ_μ, which is determined by the values of $\{w_j(\mu)\}_{j=1}^m$. The system of equations for $\{\tilde{u}_i(\mu)\}_{i=1}^{n_m}$ is then obtained by Galerkin projection using the reference reduced basis $\phi_{1,\dots,n_m}$.

For example, the already considered bilinear form corresponding to a Laplace operator reads, c.f. Equation (2.43), reads

$$\int_\Omega \nabla(\phi_j \circ \Phi_\mu) \cdot \nabla(\phi_j \circ \Phi_\mu)\, \mathrm{d}x = \int_{\Phi_\mu(\Omega)} \nabla \phi_j \cdot [D\Phi_\mu^{-1}]^{-1} [D\Phi_\mu^{-1}]^{-T} \nabla \phi_j \det D\Phi_\mu^{-1}\, \mathrm{d}y$$

where $\Phi_\mu(\Omega) = \Omega$ and $D\Phi_\mu^{-1} = \mathrm{Id} - \sum_{j=1}^m w_j(\mu)\, D^2 \xi_j^c$.

Remark 6.5. *The drawback is that these forms have to be assembled for every new parameter value, and the computational cost for this depends on the dimension of the full-order problem. This is a challenge to any projection-based model order reduction method that utilizes a parameter-dependent mapping and requires hyper-reduction techniques to solve, see Section 2.5.*

Remark 6.6. *If the parameters are time-dependent, or time is itself a parameter, the online phase will also feature an additional advection-like term:*

$$\frac{\mathrm{d}}{\mathrm{d}t} u_{\mathrm{trb}}(\mu) = \sum_{i=1}^{n_m} \left(\frac{\mathrm{d}\tilde{u}(\mu)_i}{\mathrm{d}t} \, \phi_i \circ \Phi_\mu + \tilde{u}(\mu)_i \, \frac{\mathrm{d}\Phi_\mu}{\mathrm{d}t} \cdot (\nabla \phi_i \circ \Phi_\mu) \right) \tag{6.13}$$

In the reference domain, this requires the evaluation of

$$\frac{\mathrm{d}\Phi_\mu}{\mathrm{d}t} \circ \Phi_\mu^{-1} = -[D\Phi_\mu^{-1}]^{-T} \, \frac{\mathrm{d}\Phi_\mu^{-1}}{\mathrm{d}t}. \tag{6.14}$$

Evaluating the latter expression is done using

$$\frac{\mathrm{d}\Phi_\mu^{-1}}{\mathrm{d}t} = -\sum_{j=1}^{m} \frac{\mathrm{d}w_j(\mu)}{\mathrm{d}t} \nabla \xi_j^c. \tag{6.15}$$

In summary, the proposed approach relies on snapshot remapping. The difference to other existing methods of this form is how the mappings are obtained. Our approach is data-driven and based on a POD of Monge embeddings. Other choices for parameter dependent mappings in the literature include problem-dependent parametrizations [35], polynomial expansion [97, 142], and high-fidelity piece-wise polynomial mappings [128].

6.4 Invertibility and boundary conditions

For now, assume $\Omega = [0,1]^{d \in \{2,3\}}$, the unit square or cube. Proposition 2.3 in [128] proves two sufficient conditions in order for a mapping of the form $\Phi^{-1}(y) = y - \sum_{j=1}^m w_j \nabla \xi_j^c$ to be a bijection in this case: Firstly, $\nabla \xi_j^c \cdot \hat{e}_i = 0 : 1 \leq i \leq d$ on all edges (or, if $d = 3$, faces), where $\hat{e}_{1,\dots,d}$ are normal vectors, and secondly $\det D\Phi^{-1} > 0$ in $\Omega \cup \partial\Omega$. In the case of more general mappings from Ω_2 to Ω_1, the first condition reads $\mathrm{dist}(\Phi^{-1}(y), \partial\Omega_1) = 0 \; \forall y \in \partial\Omega_2$ (Proposition 2.4 therein).

Un-regularized case

A natural question is if these conditions are met by the mappings defined in this section. Let us begin by considering the case without entropic regularization. Denote by ψ^c the optimal transport potential for the transport from $\bar{\rho} \in \mathcal{P}_{\mathrm{ac}}(\Omega_2)$ to $\rho \in \mathcal{P}_{\mathrm{ac}}(\Omega_1)$. From Section 3.3 we know that the second boundary value problem for the Monge-Ampére equation that $\varphi^* = \frac{1}{2}|\cdot|^2 - \psi^c$ solves implies that $\mathrm{id} - \nabla\psi^c$ maps $\partial\Omega_1$ into $\partial\Omega_2$ provided that $\mathrm{supp}(\bar{\rho}) = \overline{\Omega}_1$ and $\mathrm{supp}(\rho) = \overline{\Omega}_2$. Invertibility of $\mathrm{id} - \nabla\psi^c$ requires strict convexity of φ^*. This is given by the following theorem, due to Caffarelli [34].

Theorem 6.1 (Sufficient conditions for strictly convex transport maps ([47], Theorem 2.2))**.** *Let $\rho, \sigma \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ with $\mathrm{supp}(\rho) = \overline{\Omega}_1$, $\mathrm{supp}(\sigma) = \overline{\Omega}_2$, and both $\partial\Omega_1$ and $\partial\Omega_2$ are of Lebesgue measure zero. Furthermore, assume that there exists a constant $\Lambda > 0$ such that $\Lambda \leq \rho, \sigma \leq 1/\Lambda$ on the respective supports of the densities. Furthermore, assume that Ω_2 is convex. Then, if $\nabla\varphi$ is the optimal transport map between ρ and σ, φ is strictly convex. The modulus of strict convexity of φ depends only on $\Lambda, \Omega_1,$ and Ω_2.*

According to Theorem 6.1, it is enough for $\bar{\rho}, \{\rho_i\}_i$ to be bounded below and above and Ω to be convex such that the maps id $- \nabla\psi_i^c$ are invertible for all i.

A sufficient condition to enforce positivity of the Jacobian determinant

$$\det D\Phi^{-1} = \mathrm{Id} - \sum_{j=1}^{n_s} w_j D^2 \psi_j^c \tag{6.16}$$

would be that the coefficents $w_{1,\dots,n_s}$ are normalized, non-negative weights $\omega_{1,\dots,n_s}$, and $\det(\mathrm{Id} - D^2\psi_{j^*}^c) > 0, \omega_{j^*} > 0$ for at least one j^*. Again, this is the setting of Monge embedding barycenters.

By opting for linear combinations over convex ones in order to make use of the POD compression on the tangent space, we lose the guaranteed bijectivity. A similar approach is taken in [128].

Regularized case

In the entropic case, the transport maps are also gradients of convex functions (Remark 4.8). However, the entropic Monge map does not exactly solve the optimal transport boundary value problem. Since ψ^ε is defined as a Gaussian convolution of $\exp(\psi^{c,\varepsilon}/\varepsilon)\sigma$, at best we can assume that the potential is very small once $\mathrm{dist}(x, \mathrm{supp}(\sigma)) > 3\varepsilon^{1/2}$.

Because of this, the boundary condition has to be enforced in a post-processing step. Denote by $\psi^c_{\mathrm{pre-proj.}}$ denotes the output of the entropic optimal transport calculations. We opt for a global correction, by finding a potential ψ^c closest to $\psi^c_{\mathrm{pre-proj.}}$ in a weighted H^1 norm that satisifes $\nabla \cdot \psi^c \cdot \hat{n} = 0$ on $\partial\Omega$.

To be precise, we solve the system

$$\begin{aligned}\int_\Omega (\kappa^2 \nabla\psi^{c,\varepsilon} \cdot \nabla v + \psi^{c,\varepsilon} v) + \delta^{-1} \int_{\partial\Omega} (\nabla\psi^{c,\varepsilon} \cdot \hat{n}) v \\ = \int_\Omega (\kappa^2 \nabla\psi^{c,\varepsilon}_{\mathrm{pre-proj.}}(\mu_i) \cdot \nabla v + \psi^{c,\varepsilon}_{\mathrm{pre-proj.}}(\mu_i) v) \quad \forall v \in V_h\end{aligned} \tag{6.17}$$

for every $i = 1, \dots, n_s$.

There are two parameters to set: δ is a penalty term to enforce the boundary condition and is set to 10^{-9} in our numerical experiments. The value of κ determines the scale on which the function changes shape to fit the boundary condition. Since we expect the error introduced by the entropic smoothing to be of the scale $\sqrt{\varepsilon}$ and we want to limit the number of free parameters in our method, we set $\kappa^2 = \varepsilon^{-1}$.

Equation (6.17) implies that $\psi^c_\Delta := \psi^{c,\varepsilon}_{\mathrm{pre-proj.}} - \psi^{c,\varepsilon}$ solves

$$-\kappa^2 \Delta \psi^c_\delta + \psi^c_\delta = 0 \tag{6.18}$$

in Ω with the Neumann boundary condition

$$\partial_n \psi^c_\delta = \partial_n \psi^{c,\varepsilon}_{\mathrm{pre-proj.}} \tag{6.19}$$

on $\partial\Omega$. This problem is well defined, c.f. [90], Section 3.3 and in particular Theorem 3.15.

Remark 6.7. *If $\psi^{c,\varepsilon}_{\text{pre-proj.}}$ is very far from fulfilling the boundary conditions, this step can deform the potential to the point that $y \mapsto \frac{|y|^2}{2} - \psi^c(y)$ is no longer convex and the mapping no longer invertible. It is therefore crucial that ε is chosen small enough such that $\psi^{c,\varepsilon}_{\text{pre-proj.}}$ is close to ψ^c, the optimal potential of the un-regularized problem, that will fulfill the boundary condition.*

We show the effect of the H^1 projection in Figure 6.1. The projected transport potential is taken from the example presented in Section 6.8.1.

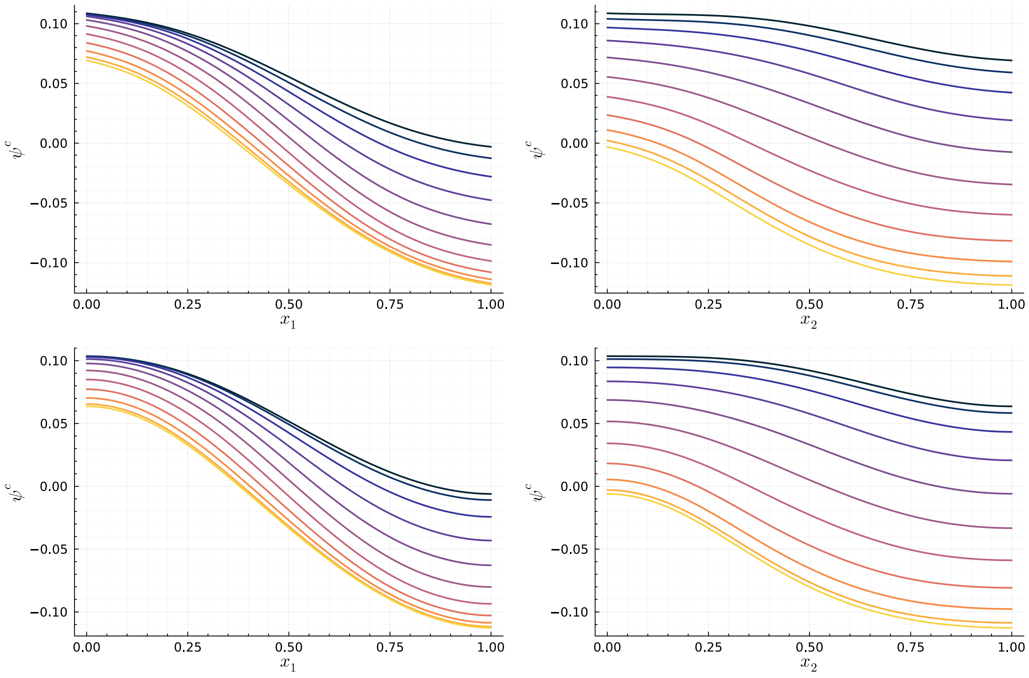

Figure 6.1: Cross-sections of ψ^c before (top row) and after (bottom row) application of the boundary projection. The depicted transport potential is taken from the example problem discussed in Section 6.8.1 and corresponds to $\mu \approx [0.348, 0.174]$. We see that the correction at the boundary is minimal. In this example, $\varepsilon = \kappa^{-2} = 10^{-2}$.

6.5 Regularity

When expressing the residual of the PPDE in the reference domain as in Equation (2.43), we require additional regularity of Φ_μ^{-1} to allow for the computation of its derivatives. A sufficient condition is $\Phi_\mu^{-1} \in \mathcal{C}^1(\Omega, \mathbb{R}^d)$. We therefore require the optimal transport potentials $\{\psi_i^c\}_{i=1}^{n_s}$ from which Φ_μ^{-1} is constructed to be $\in \mathcal{C}^2(\Omega)$.

For the un-regularized case, we have discussed sufficient conditions for this in Section 3.4. Recall from Caffarelli's regularity results (Theorem 3.5) that $\mathcal{C}^{k,\alpha}$ densities and convexity of the support of the target measure guarantee the optimal transport potential to be $\mathcal{C}^{k+2,\alpha}$.

When employing entropic regularization, the optimal potentials are smooth, as they are defined through a Gaussian convolution. The H^1 projection step is a elliptic problem and will preserve this regularity up to the boundary, with a constant depending on κ.

Vanishing densities

For practical purposes, it is also interesting what we can say about the size of $\|\Phi_\mu^{-1}\|_{L^\infty}$ and $\det \Phi_\mu^{-1}$. Consider the example from Proposition 6.1. As shown in Appendix C, the transport map reads

$$T_{\rho_{\bar{\mu}} \to \rho_\mu}(y) = 1 - \frac{1}{\mu} \sinh^{-1}\left(\frac{\sinh \mu}{\sinh \bar{\mu}} \sinh(\bar{\mu}(1-y)) \right) \tag{6.20}$$

and therefore

$$\partial_y T_{\rho_{\bar{\mu}} \to \rho_\mu}(1) \approx \tfrac{\bar{\mu}}{\mu} e^{\mu - \bar{\mu}}. \tag{6.21}$$

For typical values of $\epsilon^2 = 10^{-1}, \mu_{\min} = 20, \bar{\mu} \approx 78$, these values can be as extreme as 10^{52} or 10^{-25}, which makes this mapping unusable in practice.

In practice, the best option to prevent $\det D\Phi_\mu^{-1}$ from taking extreme values is via the bounds on the densities $\bar{\rho}$ and $\rho(\mu)$, since

$$\det D\Phi_\mu^{-1} \approx \det D^2 \varphi_{\bar{\rho} \to \rho(\mu)} = \frac{\bar{\rho}}{\rho(\mu) \circ \varphi_{\bar{\rho} \to \rho(\mu)}}. \tag{6.22}$$

The entropic smoothing for optimal transport barycenters discussed in Section 5.3.2 can provide this effect as well.

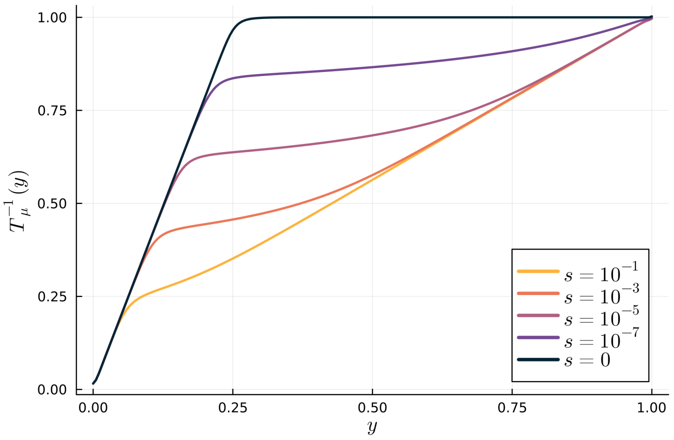

Figure 6.2: Transport maps for the problem from Proposition 6.1 with $\epsilon^2 = 10^{-1}$ and $\mu_{\min} = 20$ for different values of s with ε constant at 10^{-4}. We see that the parameter s that sets a lower bound to the densities can control the derivatives of the mapping in this example. Note that modification of T beyond the point $x^* \approx \frac{\mu}{\bar{\mu}} \approx 0.26$ does not impact the approximation result in Proposition 6.1 (Appendix C) so that the choice $s = 10^{-7}$ provides the same error bounds while keeping the derivatives of T and T^{-1} under control.

As an illustration, when using $\rho(u) = (1-s)u/\int u + s$ with $s > 0$ in Proposition 6.1, we see that the derivative of the transport map is controlled (Figure 6.2). Note in particular that the bound on the derivative is in fact much better than the obvious $(\max \rho)/(\min \sigma)$ type bounds, which would imply $\partial_y \Phi_\mu^{-1}(y) \in (s/(\mu + s), (\bar{\mu} + s)/s)$.

Mass splitting

The results for the un-regularized case are also of practical interest for the sake of guarantees as $\varepsilon \to 0$. Recall from Section 3.4 and in particular Example 3.2 that transport to non-convex target sets can lead to singularities. The entropic regularisation avoids these singularities, but at the same time does not exactly fulfil the marginal constraints. As $\varepsilon \to 0$, the entropic transport map T^ε converges to the un-regularized T, which can be discontinuous.

In Figure 6.3 and Figure 6.4 we illustrate this for the transport problem between the uniform measure on a disk and a crescent shape in $\Omega = [0,1]^2$. The entropic OT problem is solved using the setup from Section 6.8 with $\varepsilon = 10^{-2}$. Since the computational method is designed for strictly positive measures, we set the density outisde of the respective shapes to 10^{-16}.

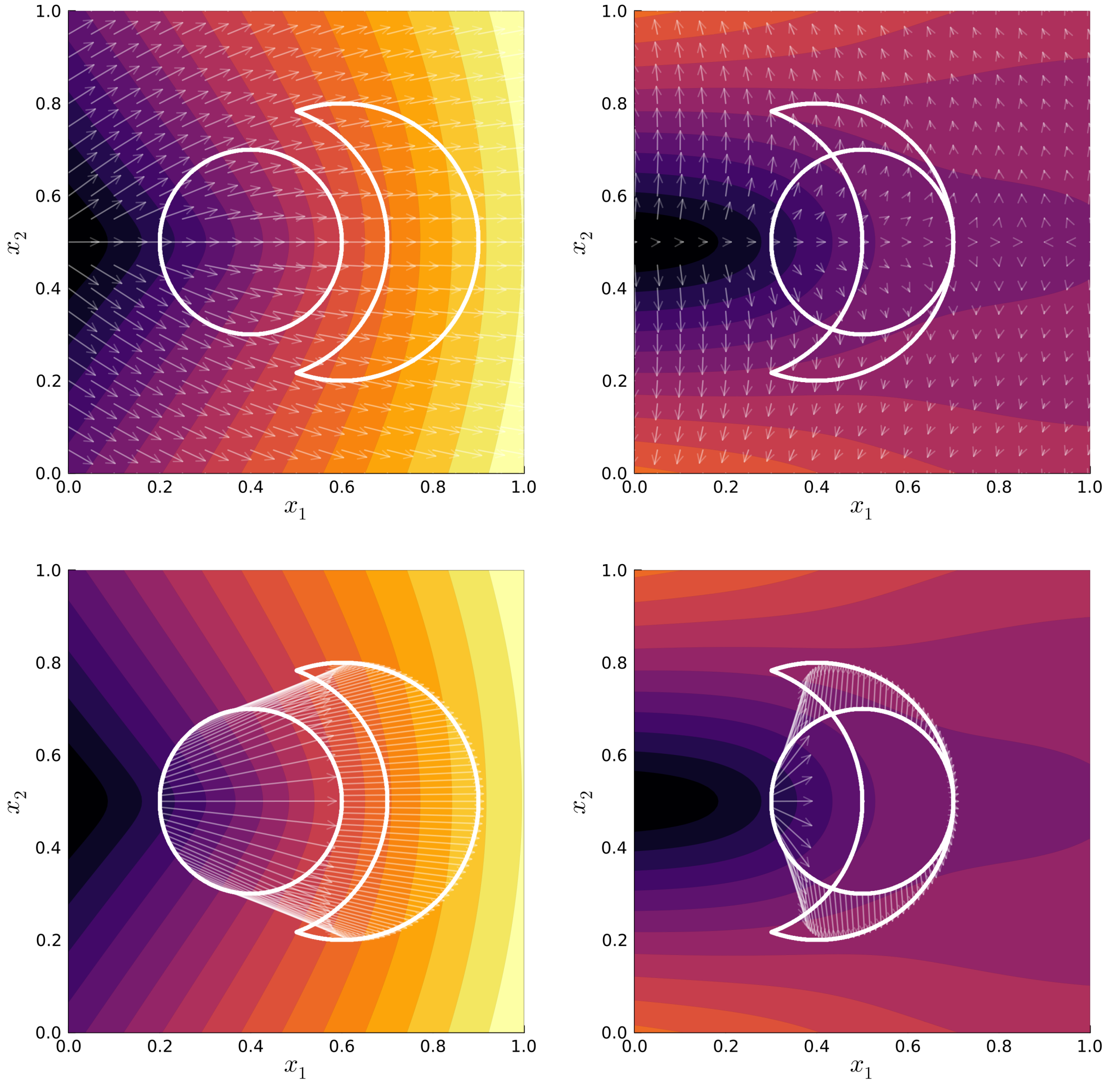

Figure 6.3: Visualization of the de-biased entropic optimal transport potential ψ^ε and the displacements $\frac{1}{4}(\mathrm{id} - \nabla\psi^\varepsilon)$ for the uniform measure on a ball and a crescent. In the first row, the displacements are re-scaled for a clearer plot. The second row shows the displacements for all points on the boundary.

We observe that the x_2-derivative of the mapping is well under control, however this comes at the price of violating the marginal constrains. As expected, the violation is of the order $\sqrt{\varepsilon} = 0.1$. We furthermore observe that the violations are more pronounced at points where the un-regularized mapping would develop discontinuities.

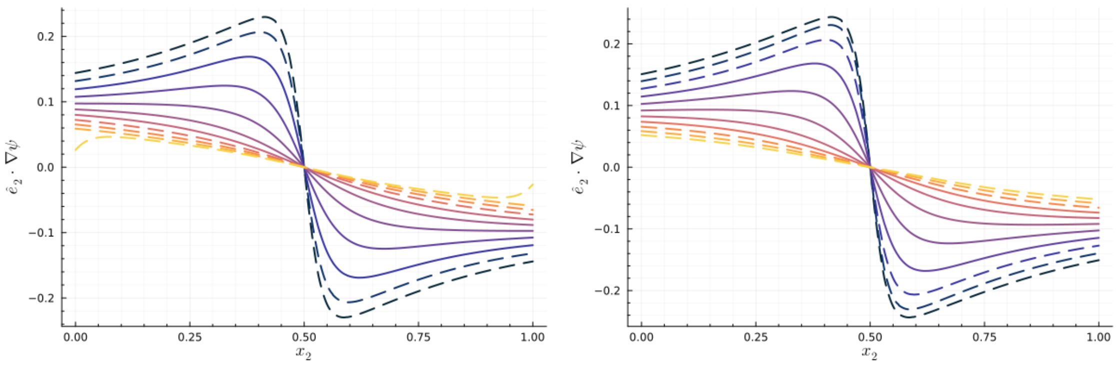

Figure 6.4: Derivatives of the de-biased entropic optimal transport potentials ψ^ε from Figure 6.3 in the x_2 direction for $x_1 \in \{0, 0.1, \dots, 1\}$ from dark to bringt. Solid lines denote those cases where the plotted cross-sections intersect the support of the ball (the transported density).

6.6 Hyper-reduction

When the online phase can be made fully independent of the size N of the high-fidelity problem, its computational cost can be reduced dramatically. For now, the assembly of linear and bilinear forms has to be done online and depends on N (c.f. Remark 6.5).

We can remedy this shortcoming using the empirical interpolation method introduced in Section 2.5 and in particular Algorithm 3. We utilize a version of EI based on a POD of the parameter-dependent forms [39, 128]. We briefly recall the method using the example of the mapped Laplace operator from Equation (2.43).

Based on the data from the training set, a collection of snapshots

$$\{[D\Phi_{\mu_i}^{-1}]^{-1}[D\Phi_{\mu_i}^{-1}]^{-T} \det D\Phi_{\mu_i}^{-1}\}_{i=1}^{n_s} =: \{K_{\mu_i}\}_{i=1}^{n_s} \tag{6.23}$$

is used to obtain a POD basis from the correlation matrix $\mathbb{C}^K$ with entries

$$\mathbb{C}_{ij}^K = \int_\Omega \operatorname{tr}(K_{\mu_i}^T K_{\mu_j})\, dy, 1 \leq i, j \leq n_s. \tag{6.24}$$

Using an energy criterion τ_{eim}, the eigenvectors $\Xi_q : 1 \leq q \leq Q$ corresponding to the Q largest eigenvalues are selected to span an approximation space. Coefficients $\theta_q(\mu)$ and functions $X_q : 1 \leq q \leq Q$ are determined such that $K_\mu \approx \sum_{q=1}^Q \theta_q(\mu) X_q$ for all μ. The way the interpolation points and functions are selected guarantees that the matrix $B \in \mathbb{R}^{Q \times Q} : B_{q'q} = X_{q'}(y_q^{\text{eim}})$ is lower-triangular with unit diagonal, so the interpolation problem is well-defined and quickly (i.e. in $\mathcal{O}(Q^2)$ time) solved.

Online, K_μ is evaluated at the points $\{y_q^{\text{eim}}\}_{q=1}^Q$ and the interpolation problem $K_\mu(y_q^{\text{eim}}) = \sum_{q'=1}^Q \theta_{q'}(\mu) X_{q'}(y_q^{\text{eim}}) : 1 \leq q \leq Q$ is solved to obtain $\{\theta_q(\mu)\}_{q=1}^Q$. The full form is approximated using

$$\int_\Omega \nabla\phi_j \cdot K_\mu \nabla\phi_j \,\mathrm{d}y \approx \sum_{q=1}^Q \theta_q(\mu) \int_\Omega \nabla\phi_j \cdot X_{q,ij} \nabla\phi_j \,\mathrm{d}y, \tag{6.25}$$

where the integral defines a $Q \times n_m \times n_m$ tensor that can be pre-computed offline. As a result, no integration in the high-fidelity space has to be performed in the online phase.

The online cost of the EIM procedure consists of Q evaluations of K_μ, the $\mathcal{O}(Q^2)$ interpolation problem, and a $\mathcal{O}(Qn_m^2)$ tensor contraction. Importantly, it does not depend on N.

6.7 Comparison with other works

In this section, we contrast the proposed method to three approaches that have been proposed in the past and relate to this work. Naturally, any of the registration methods cited so far [112, 128, 97, 35, 142, 72] would be a candidate for comparison.

6.7.1 Optimization-based registration

In [128], the author presents an optimization-based registration method for PPDE problems. Given a set of snapshots $\{u(\mu_i)\}_{i=1}^{n_s}$ with $\{\mu_i\}_{i=1}^{n_s} \subset \mathcal{A}$, it returns a parameter-dependent bijective mapping $\Phi : \mathcal{A} \times \Omega \to \Omega$.

Objective function

The mappings are approximated as

$$\Phi_\mu^{\text{hf}}(y) := \sum_{j=1}^{m^{\text{hf}}} w^{\text{hf}}(\mu)_j \chi_j^{\text{hf}}(y). \tag{6.26}$$

As in Equation (6.7) , $\{w^{\text{hf}}(\mu)_j\}_{j=1}^{m^{\text{hf}}}$ are parameter-dependent coefficients, while $\{\chi^{\text{hf}}\}_{i=1}^{m^{\text{hf}}}$ are elements of a general approximation space such as Legendre polynomials or Fourier expansions. The mappings are constructed such that

$$\operatorname{Prox}(\mu, \Phi^{\text{hf}}) := \left\| u(\mu) \circ \left(\text{id} + \sum_{j=1}^{m^{\text{hf}}} w_j^{\text{hf}} \chi_j^{\text{hf}} \right) - \bar{u} \right\|_{L^2(\Omega)}^2 \tag{6.27}$$

is small, given a reference $\bar{u}$. Besides this proximity measure, the optimization penalizes the H^2 semi-norm of the mappings and enforces constraints to keep the Jacobian of the mappings strictly positive. In particular, the complete minimization problem reads

$$\min_{\{w_j^{\mathrm{hf}}\}_j \in \mathbb{R}^{m^{\mathrm{hf}}}} \Big(\mathrm{Prox}(\mu, \Phi^{\mathrm{hf}}) + \alpha \|\Phi^{\mathrm{hf}}\|^2_{\dot{H}^2(\Omega)} \Big)$$

$$\text{subject to } \left\| \exp \frac{\epsilon - \det D\Phi^{\mathrm{hf}}}{C} + \exp \frac{\det D\Phi^{\mathrm{hf}} - 1/\epsilon}{C} \right\|_{L^1(\Omega)} \leq \delta. \quad (6.28)$$

The hyper-parameters ϵ, C, and δ are set to $0.1, 0.025\epsilon$, and $|\Omega|$, respectively in [128]. The value of α has to be chosen depending on m^{hf} and can vary substantially - between 10^{-10} and 10^1. The constraint weakly enforces $\epsilon < \det D\Phi^{\mathrm{hf}} < 1/\epsilon$.

Reduction

To guarantee a sufficiently rich set of mappings to optimize over, m^{hf} has to be rather large, which can be restrictive when evaluating mappings in the online phase. Consequently, the authors also opt for a POD approach, reducing the number of mapping terms to m based on an eigenvalue decomposition of the matrix $\mathbb{C}^w \in \mathbb{R}^{m^{\mathrm{hf}} \times m^{\mathrm{hf}}}$ with elements

$$\mathbb{C}^w_{ij} = w^{\mathrm{hf}}(\mu)_i \cdot w^{\mathrm{hf}}(\mu)_j. \quad (6.29)$$

The POD approximation leads to

$$\sum_{j=1}^{m^{\mathrm{hf}}} w^{\mathrm{hf}}(\mu)_j \chi_j^{\mathrm{hf}} \approx \sum_{j=1}^{m} w(\mu)_j \chi_j. \quad (6.30)$$

This method is similar to the one we propose in the present work. Note that also in this case, it cannot be guaranteed that this approximate mapping is invertible, even if the high fidelity one is. However, the method proved stable in the numerical test cases considered.

Connection to optimal transport maps

The mappings Φ_μ^{hf} correspond to the transport maps $y \mapsto y - \nabla\psi_\mu^c(y)$ in the present work. Note that similar constraints can be enforced in this case: The upper and lower bounds on $\rho(\mu)$ control $\det D\Phi_\mu$ through the Monge-Ampére equation (3.24). The proximity measure is related to the condition $(\mathrm{id} - \psi_i^c)_\sharp \bar{\rho} = \rho_i$. From the discussion following Definition 4.3, we expect this to be satisfied to order ε. Lastly, the transport cost $\|\nabla\psi_\mu^c\|_{L^2(\bar{\rho})}$ controls the H^1 semi-norm of ψ_i^c.

Remark 6.8. *Note that control over $\nabla\psi^c$ does not guarantee anything for the higher order derivatives of ψ^c by itself. Consider the simple example of $\rho(x) = \mathbb{1}_{[0,1]}(x)$ and $\sigma(y) = 2y\,\mathbb{1}_{[0,1]}(y)$. We find $\partial_x \psi_{\rho\to\sigma}(x) = x - \sqrt{x}$, so $\|\partial_x\psi\|^2_{L^2(\rho)} = 1/30$, and $\|\partial_x^2\psi\|^2_{L^2(\rho)} = +\infty$. We assume that more can be said for the case where $\Lambda \leq \rho, \sigma \leq 1/\Lambda$, but these are non-trivial questions even when we know that ψ is smooth (c.f. Section 3.4).*

One classical result in this direction is given in [36]: Let $\Omega \subset \mathbb{R}^d$ open with $d \leq 5$ contain the ball $B(0,r)$. Furthermore, assume $\varphi \in \mathcal{C}^5(\Omega)$ solves $\det D^2\varphi = 1$ on Ω and that $D^2\varphi(0) = \mathrm{Id}$. Then, $\sum_{i,j,k=1}^d (\partial_{x_i}\partial_{x_j}\partial_{x_k}\varphi)^2(0) \leq 4M_d/r^2$, where M_d are universal constants with $M_2 \leq 4$ and $M_3 < 2660$.

6.7.2 Point set registration

In the later work [129], the authors employ a similar optimization-based registration method for point registration applications. In this case, the proximity measure is based on a number of pairs $(x^{(i)}, y^{(j)}) : i = 1, \dots, n_p$ and the mapping has to satisfy $\Phi(x_i) \approx y_i \; \forall i$, with the discrepancy measured either as the average error $\sum_i |\Phi(x_i) - y_i|^2$ or weakly enforced as $\max_i |\Phi(x^{(i)}) - y^{(j)}|_\infty \leq \delta$. Furthermore, additional terms enforce boundedness of the Jacobian determinant of Φ or the anisotropy of the mesh.

Restriction to gradient mappings in the optimization

The author compares the best mapping obtained by the method in two cases: In the first case, Φ is obtained through optimization on general polynomial functions up to a certain degree on Ω valued in $\mathbb{R}^2$. In the second case, Φ is restricted to be of the form $\Phi(x) = x + \nabla\varphi$ for a real-valued polynomial function φ. In both cases, $\Phi(\partial\Omega) = \partial\Omega$ for all candidate mappings. The polynomial degree in the two cases are chosen such that the trial spaces have comparable dimension.

It is observed (c.f. Figure 5 in [129]) that while the mapping is of similar quality for most deformations, the optimization returns worse or even inadmissible mappings when restricted to gradient functions in the case of large deformations. This poses questions about the quality of optimal transport maps for the purposes of registration.

Entropic transport maps for point set registration

In the example considered where $n_p = 3$ points are matched, the optimal transport problem is clearly trivial. However, the optimal transport potential is only defined at the point values $\{x_i\}_{i=1}^{n_p}$. While entropic regularization defines $\psi^{c,\varepsilon}$ for all values in Ω (or in $\mathbb{R}^d$, even), this extension is not necessarily a good map. By substituting the formula for two densities of the form $\rho = \sum_i \hat\rho_i \delta_{x^{(i)}}, \sigma = \sum_i \hat\sigma_j \delta_{y^{(j)}}$ into Equation (4.9), we obtain

$$\exp\left(-\frac{\psi^{c,\varepsilon}(y)}{\varepsilon}\right) \\ = \sum_i \exp\left(\frac{-|x^{(i)} - y|^2}{2\varepsilon}\right) \hat\rho_i \left(\sum_j \exp\left(\frac{-|x^{(i)} - y^{(j)}|^2}{2\varepsilon} + \frac{\psi^{c,\varepsilon}(y_j)}{\varepsilon}\right) \hat\sigma_j\right)^{-1}. \tag{6.31}$$

Therefore, as $\varepsilon \to 0$,

$$y - \nabla\psi^{c,\varepsilon}(y) = T^\varepsilon_{\sigma\to\rho}(y) \approx x^{(i^*)}, \text{ where } i^* := \operatorname*{arg\,min}_{1 \leq i \leq n_p} |x^{(i)} - y|^2. \tag{6.32}$$

The proposed method is not designed for point registration approaches, as it requires ρ to admit a density that is bounded from below in the entire domain. That said, if ε is chosen large enough and the points to be registered are approximated by narrow (e.g.) Gaussians, one can apply it to the problem in [129]. We let $\varepsilon = 10^{-2}$ and discretize the domain $\Omega = [0, 1]^2$ using a uniform 96 × 96 grid of quadrilaterals. The transported densities are approximated as Gaussians with standard deviation equal to 7×10^{-3} centered at the registration points:

$$\{x^{(i)}\}_{i=1}^3 = \{(1/2, 1/2), (1/4, 1/4), (3/4, 1/4)\} \tag{6.33}$$

and

$$\{y^{(j)}\}_{j=1}^{3} = \{(1/4, 3/4), (1/16, 1/16), (1/2, 1/4)\} \tag{6.34}$$

Since both $\rho = \frac{1}{3}\sum_i \delta_{x^{(i)}}$ and $\sigma = \frac{1}{3}\sum_j \delta_{y^j}$ are effectively zero throughout most parts of the domain, the OT boundary condition $\Phi(\mathrm{supp}(\rho)) = \mathrm{supp}(\sigma)$ does not resemble $\Phi(\Omega) = \Omega$. We therefore apply the proposed H^1 projection to enforce it.

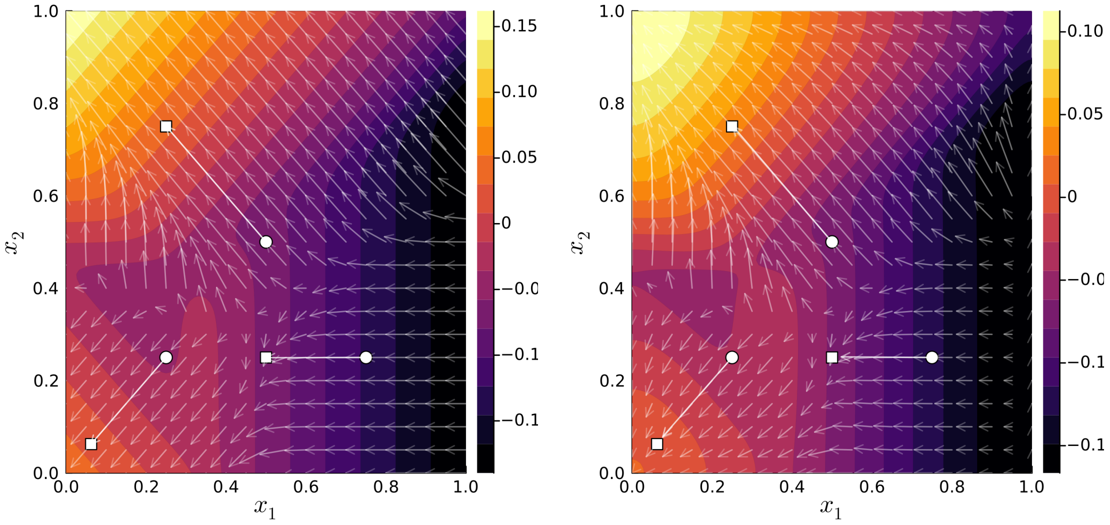

Figure 6.5: Visualization of the entropic optimal transport potential ψ^ε and the displacements $\frac{1}{4}(\mathrm{id} - \nabla\psi^\varepsilon)$ for the point registration example before (left) and after (right) application of the boundary condition projection from Equation (6.17). The displacements are re-scaled for a clearer plot, at the points $\{x_i\}_i$, the longer arrows show the true displacement $\mathrm{id} - \nabla\psi^\varepsilon$.

Figure 6.5 displays the contour lines of the entropic OT potential, extended to all $x \in \Omega$ through the Schrödinger equations Equation (4.9). Note that the boundary condition is not met by a significant margin. After projection, it holds.

However, while we find $\max_i \min_j |x^{(i)} - \nabla\psi^\varepsilon(x^{(i)}) - y^{(j)}| = 6.9 \times 10^{-4}$ before the boundary conditions are enforced, this error grows to $\max_i \min_j |x^{(i)} - \nabla\psi^\varepsilon(x^{(i)}) - y^{(j)}| = 2.3 \times 10^{-2}$ in the right picture. For the inverse mapping, since $y^{(2)}$ lies very close to the domain boundary, this effect is even more pronounced, and $\max_j \min_i |y^{(j)} - \nabla\psi^{c,\varepsilon}(y^{(j)}) - x^{(i)}|$ grows from 4.4×10^{-4} to 1.4×10^{-1} by applying the boundary condition projection.

This example was discussed to illustrate the extension of the transport potential defined at point values $\{x^{(i)}\}_i$ to the entire domain Ω. If the transport map is to respect the boundary condition $\Phi(\Omega) = \Omega$, then we necessarily need to have mass near $\partial\Omega$.

How restrictive are gradient mappings?

It is an interesting result that the performance of the optimization-based method degrades when restricting the trial space to gradient mappings. In the case of mass transport of continuous density, the restriction of transport maps S to the form $S = \nabla\varphi$ is the basis of the regularity theory of the Monge-Ampére equation. In

[47], this is compared to the situation of a divergence-free vector field $v : \nabla \cdot v = 0$, which can be arbitrarily irregular, and a divergence-free gradient field $\varphi : \nabla \cdot \nabla \varphi = 0$, which is as regular as the boundary conditions permit.

It is also worth noting that the gradient maps perform as well as their superset of polynomial vector-valued maps as long as the deformations are not large. To be more precise, the points $\{y^{(i)}\}_i$ are moved along the lines $x^{(i)} + t(y^{(i)} - x^{(i)})$ for $t \in [0.1, 1]$. For all $t \leq 0.8$, the gradient maps yield good results. The case shown in Figure 6.5 is $t = 1$.

Gradient maps will certainly fail in an optimization-based approach if the registration problem corresponds to a case where the optimal transport map is discontinuous such as Caffarelli's Example 3.2. In this case, the optimization will not be able to fulfil the regularity requirements since to do so, it would have to violate the cyclic monotonicity. However, the latter is enforced by construction of the trial space.

6.7.3 Advection modes

The registration method presented in [72] uses OT techniques to build the mapping Φ_μ. Compared to the present work, there are three main differences: Firstly, the mapping of the snapshots is done in the sense of push-forward measures, that is

$$\rho(\mu) \approx (\bar{\rho} \circ \Phi_\mu + r(\mu) \circ \Phi_\mu) \det D\Phi_\mu \tag{6.35}$$

where $r(\mu)$ is a residual defined in the reference configuration. Note the presence of the Jacobian determinant of the mapping when compared to our formulation for u_{trb} in (6.7). As in the present work and as in [128, 131] the residual is approximated by linear combinations of POD modes. The mapping is approximated as a linear combination of *advection modes*.

These modes are obtained using the distance matrix $\mathbb{D}$ with entries

$$\mathbb{D}_{ij} := W_2(\rho(\mu_i), \rho(\mu_j))^2 \quad \forall i, j \in 1, \dots, n_s \tag{6.36}$$

Recall that if one were to replace W_2 by $\| \cdot \|_2^2$, one could reconstruct the positions of $\rho(\mu_1), \dots, \rho(\mu_{n_s}) \in \mathbb{R}^N$ up to rotations and translations of the entire set solely from the relative distance information contained in $\mathbb{D}$ by performing a singular value decomposition of the matrix $\mathbb{B} := -\frac{1}{2}\mathbb{J}\mathbb{D}\mathbb{J}$ with $\mathbb{J} = \text{Id} - \frac{1}{n_s}\mathbb{1}\mathbb{1}^T$.

Furthermore, retaining only the m largest singular values, one can obtain the best approximate positions in an m-dimensional space. We refer again to the comprehensive review of Euclidean distance matrix methods given in [52].

For the W_2 distance, however, we know that $\mathbb{B}$ is not positive semi-definite. By omitting the negative singular values, the authors obtain approximations of the positions of $\rho(\mu_1), \dots, \rho(\mu_{n_s}$ in a low-dimensional space. By inspection, they find parametrizations of $\mu \mapsto \Phi_\mu$. For example, if the reconstructed positions lie approximately on a line, then the dominant advection mode is set as the transport map connecting the solution at its beginning to the one at the end.

The definition $u_{\text{trb}}(\mu) = \sum_{i=1}^{n_m} \tilde{u}_i(\mu)\, \phi_i \circ \Phi_\mu$ used in our work corresponds to the push-forward of a function using Φ_μ^{-1}. The same operation is used to obtain the reference reduced basis. Especially in the case where u itself is used to construct the transport mappings, one could also use the push-forward as applied on a density. In a sense, this is a natural choice, since the transport mappings are constructed to fulfil the push-forward relation for a density. Up to numerical inaccuracies, it holds on the training set that

$$\bar{\rho} = \rho(\mu_i) \circ (\mathrm{id} - \nabla\psi^c(\mu_i)) \det\big(\mathrm{Id} - D^2\psi^c(\mu_i)\big) \quad \forall i = 1, \dots, n_s, \tag{6.37}$$

and therefore the only reference reduced basis function is expected to be $\bar{\rho}/\|\bar{\rho}\|_{L^2}$.

We chose to not follow this approach for two reasons: First, relation (6.37) only holds if the transport mappings are constructed directly from u itself and not another derived quantity. This naturally limits the range of application cases.

Second, substituting the push-forward relation for a density into a PDE residual requires evaluating derivatives of $\det(\mathrm{Id} - D^2\psi^c(\mu_i))$, which requires even more regularity of ψ^c, higher order basis functions in the discretization., and a higher order numerical quadrature in the high-fidelity problem.

6.8 Numerical examples

We will demonstrate the proposed method and the impact of some of the hyperparameters on two test cases. For finite element calculations, we rely on the `Gridap.jl` library[1] [136, 11], while `GaussianProcesses.jl`[2] is used for the computation of the Gaussian processes. The computational optimal transport routines used are available in the package `WassersteinDictionaries.jl`[3]. The code to reproduce the examples in this section is published in the package `OptimalMappings.jl`[4].

6.8.1 Poisson's equation with moving source

Let $\Omega = [0, 1]^2$, discretized by a 64×64 grid of quadrilateral cells. For V_h, we chose H^1_0-conforming Legendre basis functions of order $p = 3$, which leads to $N = 36481$ degrees of freedom. The grid size is denoted with h. The equation we solve is

$$\Delta u(x; \mu) = f(x; \mu) : x \in \Omega, u(x; \mu) = 0 : x \in \partial\Omega, \tag{6.38}$$

where f is a narrow Gaussian with variance $\mathrm{var} = 10^{-3}$ and mean $(\frac{1}{2}, \frac{1}{2}) + \mu$, $\mu \in [-\frac{7}{20}, \frac{7}{20}]^2 = \mathcal{A}$.

Training set and parameter choices

To construct the mappings and reduced basis we draw $n_s = 100$ samples of μ uniformly from $\mathcal{A}$. The solutions to this equation are not probability densities, so we define *rho* either as

$$\rho(u)(\mu) = \frac{u(\mu)^2}{\int u(\mu)^2} \quad \text{or} \quad \rho(u)(\mu) = f(\mu) \tag{6.39}$$

to calculate the transport mappings. These computations, which rely on collocation, are performed on a finer 192×192 grid of quadrilaterals. The iterative OT solver

[1] https://github.com/gridap/Gridap.jl

[2] https://github.com/STOR-i/GaussianProcesses.jl

[3] https://github.com/JuliaRCM/WassersteinDictionaries.jl

[4] https://github.com/ToBlick/OptimalMappings.jl

is set to stop when the l_1 error of the marginal constraint reaches 10^{-3} or at the maximum number of iterations of $\lceil 10/\varepsilon \rceil$.

As reference density $\bar{\rho}$ we chose the optimal transport barycenter of the training snapshot densities.

We emphasize again that this choice is not crucial for the proposed method. Any reference density, even $\bar{\rho} \equiv |\Omega|^{-1}$, can be used. However, when $\bar{\rho}$ is an average of the data $\{\rho(\mu_i)\}_{i=1}^{n_s}$ in a meaningful sense (which the barycenter is in these cases), we expect the transport maps to have a simpler structure and require fewer modes m to approximate. In particular, the choice of the barycenter means that the transport maps are much closer to shifts and scalings compared to using for example $\bar{\rho} \equiv |\Omega|^{-1}$. This is the setting where working on the tangent space, as we do in this method, works best.

We employ an annealing strategy as described in Section 4.2, initializing the regularization parameter to one and then halving it at every iteration until we reach ε. To invert the mapping at the last step, we use the c-transform with $\varepsilon_{\text{fine}} = h^2/10$.

Given a threshold $\tau \in (0, 1)$, m is chosen such that $1 - \mathcal{E}(m; \lambda) < \tau$, where

$$\mathcal{E}(m; \lambda) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{\operatorname{rank} \mathbb{C}} \lambda_j}. \tag{6.40}$$

We set $\tau_{\text{eim}} = 0.1\tau$ to account for the difficulty in approximating the moving source term.

The Gaussian process we use is taken as-is from the reference package. In particular, we select a zero mean function, a squared exponential kernel with characteristic length and standard deviation set to one (the default settings). The log standard deviation of observation noise is set to -6. These parameters have not been optimized.

Choice of density $\rho(u)(\mu) \propto u(\mu)^2$

Since $\rho(\mu)$ is positive on all of Ω in this case, we use debiased calculations using S_ε. The transport maps are now given by the debiased potentials as defined in Equation (4.55). We see that debiasing improves the performance of the method, both by increasing the accuracy and by reducing the number of approximation functions n_m and Q, in Table 6.2.

Figure 6.6 displays the eigenvalue decay for the correlation matrices of snapshots $\mathbb{C}^u$, transported snapshots $\mathbb{C}^{\Phi_* u}$, and Monge embeddings $\mathbb{C}^\psi$. As expected, the eigenvalues of the mapped snapshots are indicative of a much faster n-width decay of $\Phi_\mu(\mathcal{M})$ compared to that of $\mathcal{M}$.

Transport modes

Figure 6.7 shows the first four transport modes $\xi^c_{1,\dots,4}$ and the Gaussian process approximations of $\mu \mapsto w_j(\mu)$ for $j = 1, \dots, 4$, the transport mode coefficients used in the mapping $\Phi_\mu^{-1} = \mathrm{id} - \nabla \sum_{j=1}^{m} w_j(\mu)\xi^c_j$. The first two modes are essentially translations, the third mode is a contraction (or expansion, depending on the sign of its coefficient), and the fourth mode is a contraction along one diagonal and an expansion along the other.

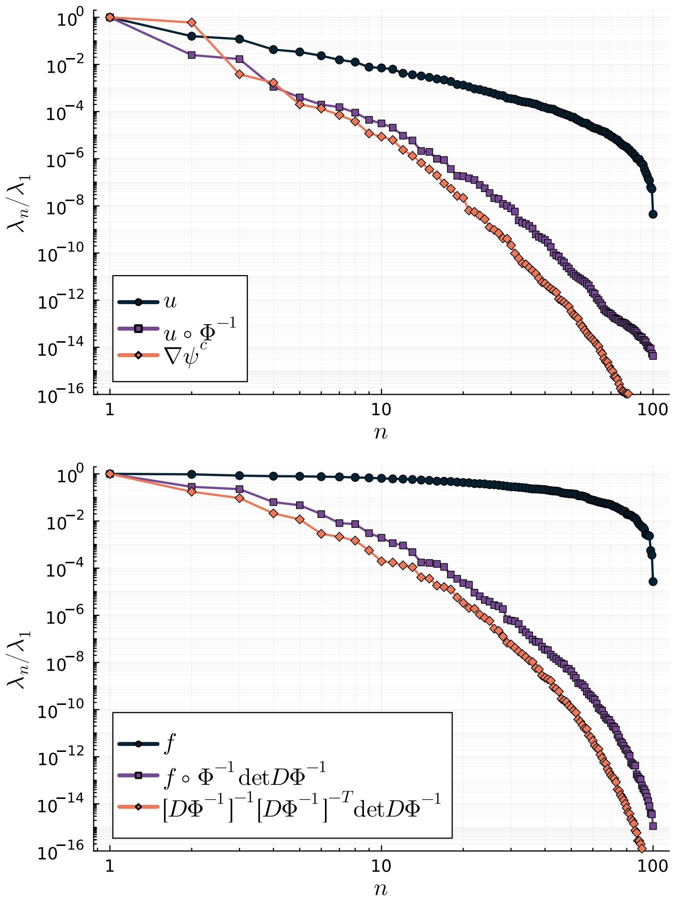

Figure 6.6: Top: Eigenvalues of the correlation matrices $\mathbb{C}^u$, $\mathbb{C}^{\Phi_* u}$, and $\mathbb{C}^\psi$. Bottom: Eigenvalues of the correlation matrices $\mathbb{C}^f$, $\mathbb{C}^K$, and $\mathbb{C}^{\Phi_* f}$ used in the hyper-reduction. $\varepsilon = 10^{-2}, \tau = 10^{-4}, \rho(u)(\mu) \propto u(\mu)^2$, and debiasing are used.

As indicated by the very fast eigenvalue decay of the correlation matrix $\mathbb{C}^\psi$, transport mappings can be approximated accurately as a linear combination of only a few transport modes.

Errors

Next, we compare the error in the solution of the PPDE for the entire online phase. This includes approximating the mapping with transport modes, obtaining the coefficients of the transport modes with a Gaussian process, solving the PPDE in the reference domain as in (2.43), and mapping the solution back to the physical domain as in (6.7).

Average and maximum errors are calculated for a test set using $n_t = 50$ samples from $\mathcal{A}$. The results are compared to the classical POD method without registration, i.e. the $m = 0$ case.

The hyper-reduction uses EI to evaluate the mapped Laplacian as described in Section 2.5 as well as the right-hand-side term $\int_\Omega \phi_i f_\mu \det D\Phi_\mu^{-1}\, dy \ \forall i = 1, \dots, n_m$. The number of interpolation functions are denoted Q_K and Q_f, respectively.

Values for the cases of $m = 0$ using hyper-reduction are not given, since the set $\{f(\mu)\}_{\mu \in \mathcal{A}}$ shows extremely slow n-width decay. In contrast, the n-width decay of

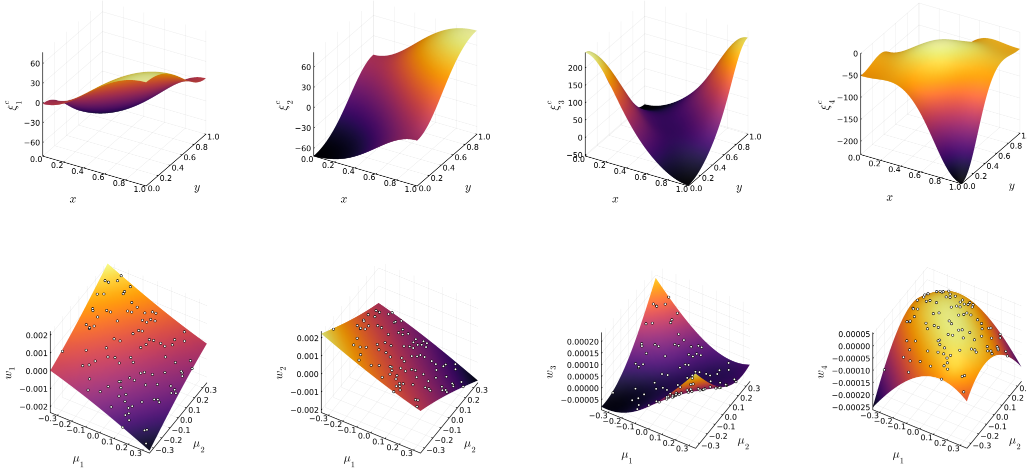

Figure 6.7: First row: first four transport modes with an added constant such that $\xi^c_{1,\ldots,4}(1/2,1/2) = 0$. Second row: Gaussian process approximation of the functions $w_{1,\ldots,4}(\mu)$. Values used to construct the basis are marked in white. The parameters chosen are as in Fig. 6.6.

$\{f(\mu) \circ \Phi_\mu^{-1} \det D\Phi_\mu^{-1}\}_{\mu \in \mathcal{A}}$ allows the use of EI albeit with large values of Q_f.

We also report the relative error of the H^1 semi-norm of $u(\mu)$, i.e. the error in the energy of the solution for an electrostatic problem. We choose this as an example of a quantity of interest that can be computed in the reference domain.

$\tau(\mathcal{E})$	n	m	Q_K	Q_f	relative L^2 error of $u(\mu)$		relative H^1 error of $u(\mu)$		relative error of $\|u(\mu)\|_{\dot{H}^1}$	
					avg	max	avg	max	avg	max
10^{-3}	41	0	-	-	7.87×10^{-2}	3.32×10^{-1}	2.99×10^{-1}	7.08×10^{-1}	1.09×10^{-1}	5.07×10^{-1}
	5	4	-	-	4.48×10^{-2}	1.27×10^{-1}	9.80×10^{-2}	2.16×10^{-1}	1.11×10^{-2}	4.67×10^{-2}
	5	4	12	19	4.94×10^{-2}	1.25×10^{-1}	1.02×10^{-1}	2.20×10^{-1}	2.12×10^{-2}	7.13×10^{-2}
10^{-4}	64	0	-	-	5.85×10^{-2}	3.19×10^{-1}	2.28×10^{-1}	6.93×10^{-1}	7.75×10^{-2}	4.85×10^{-1}
	9	6	-	-	1.53×10^{-2}	4.04×10^{-2}	4.27×10^{-2}	1.01×10^{-1}	1.71×10^{-3}	8.20×10^{-3}
	9	6	19	24	1.67×10^{-2}	5.10×10^{-2}	4.42×10^{-2}	1.08×10^{-1}	8.17×10^{-3}	4.70×10^{-2}
10^{-5}	82	0	-	-	5.10×10^{-2}	2.99×10^{-1}	2.02×10^{-1}	6.71×10^{-1}	6.68×10^{-2}	4.55×10^{-1}
	11	10	-	-	7.23×10^{-3}	2.36×10^{-2}	2.79×10^{-2}	6.20×10^{-2}	5.43×10^{-4}	2.32×10^{-3}
	11	10	28	30	9.88×10^{-3}	5.23×10^{-2}	3.02×10^{-2}	9.15×10^{-2}	6.26×10^{-3}	5.31×10^{-2}

Table 6.1: PPDE solution errors in the test set as a function of the retained eigenvalue energy for $\varepsilon = 10^{-2}, \rho(\mu) \propto u(\mu)^2$, and debiased calculations.

debiasing	n	m	Q_K	Q_f	relative L^2 error of $u(\mu)$		relative H^1 error of $u(\mu)$		relative error of $\|u(\mu)\|_{\dot{H}^1}$	
					avg	max	avg	max	avg	max
yes	9	6	-	-	1.53×10^{-2}	4.04×10^{-2}	4.27×10^{-2}	1.01×10^{-1}	1.71×10^{-3}	8.20×10^{-3}
	9	6	19	24	1.67×10^{-2}	5.10×10^{-2}	4.42×10^{-2}	1.08×10^{-1}	8.17×10^{-3}	4.70×10^{-2}
no	10	7	-	-	1.52×10^{-2}	5.42×10^{-2}	4.89×10^{-2}	1.02×10^{-1}	2.61×10^{-3}	9.97×10^{-3}
	10	7	15	35	1.97×10^{-2}	8.00×10^{-2}	5.19×10^{-2}	1.18×10^{-1}	1.22×10^{-2}	5.27×10^{-2}

Table 6.2: PPDE solution errors in the test set with and without debiasing, using $\varepsilon = 10^{-2}, \tau = 10^{-4}, \rho(\mu) \propto u(\mu)^2$.

In order to determine how much of the overall error of the method is made when inverting Φ_μ^{-1}, we compare $u_{\text{trb}} \circ \Phi_\mu^{-1}$ to $u \circ \Phi_\mu^{-1}$ in the case $\tau = 10^{-5}$ when using

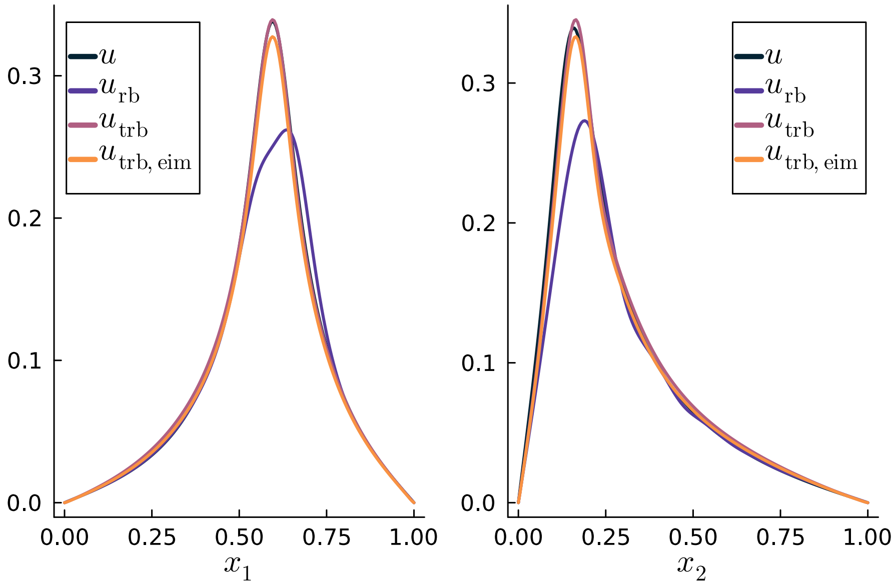

Figure 6.8: Cross-section of $u(mu)$ and its approximation by classical POD, as well as the proposed method with and without the use hyper-reduction. Depicted is that value of μ that leads to the largest H^1 error of u_{trb} in the test set.

hyper-reduction. Calculating the approximation error in the reference domain in this way, we find average and maximum L^2 errors of 9.31×10^{-3} and 5.62×10^{-2}. The average and maximum H^1 errors are 2.13×10^{-2} and 8.61×10^{-2}.

Using no hyper-reduction, these errors decrease to 5.04×10^{-3}, 2.57×10^{-2}, 1.47×10^{-2}, and 4.68×10^{-2}, respectively.

We conclude that the approximation error in the reference domain dominates in the overall error of the method. The inversion of the registration mapping using an entropic approximation of the c-transform with $\varepsilon_{\mathrm{fine}}$ proves both fast and sufficiently accurate.

Choice of density $\rho(u)(\mu) = f(\mu)$

This case is added here to see how the effect of entropic smoothing can be used to apply the method even when ρ takes very small values in Ω. In this case, we do not use debiased potentials and barycenters. Indeed, when using debiased quantities, the performance of the method is heavily degraded with this choice of ρ.

Recall that $\{f(\mu_i)\}_{i=1}^{n_s}$ consists of narrow Gaussians that differ from one another by translation. As expected, there are only three eigenvalues of $\mathbb{C}^{\psi}$ different from machine zero in when $\rho = f$. The corresponding transport modes are two translations and one scaling mode (i.e. $y \mapsto \text{const.} \cdot y^2$), which is a consequence of the entropic smoothing. The eigenvalue decay of $\mathbb{C}^K$ and $\mathbb{C}^{\Phi_* f}$ is also improved (Fig. 6.9).

Errors

Approximation errors for this case are shown in Table 6.3 and are comparable to the case $\rho(u) \propto u^2$ even at $m = 3$. However, note that the online cost is mostly

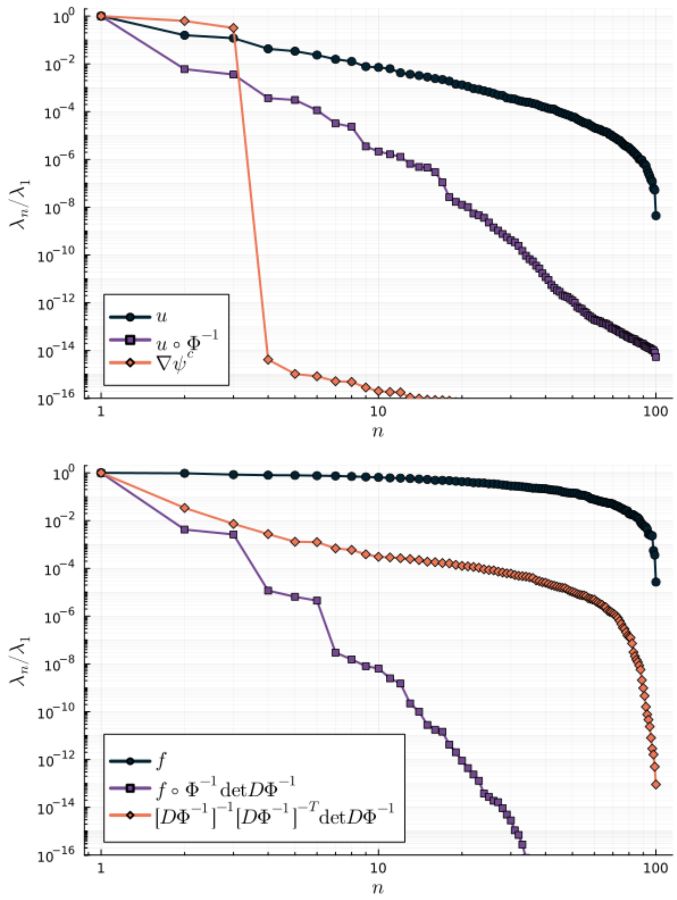

Figure 6.9: Top: Eigenvalues of the correlation matrices $\mathbb{C}^u$, $\mathbb{C}^{\Phi_* u}$, and $\mathbb{C}^{\psi}$ for the case $\rho(\mu) = f(\mu)$. Bottom: Eigenvalues of the correlation matrices $\mathbb{C}^f$, $\mathbb{C}^K$, and $\mathbb{C}^{\Phi_* f}$ used in the hyper-reduction. No debiasing, $\varepsilon = 10^{-2}, \tau = 10^{-4}, \rho(\mu) = f(\mu)$.

dependent on n, not m, see Section 6.8.1.

$\tau(\mathcal{E})$	n	m	Q_K	Q_f	relative L^2 error of $u(\mu)$		relative H^1 error of $u(\mu)$		relative error of $\Vert u(\mu)\Vert_{\dot{H}^1}$	
					avg	max	avg	max	avg	max
10^{-3}	41	0	-	-	7.87×10^{-2}	3.32×10^{-1}	2.99×10^{-1}	7.08×10^{-1}	1.09×10^{-1}	5.07×10^{-1}
	3	3	-	-	7.46×10^{-2}	2.03×10^{-1}	1.33×10^{-1}	2.68×10^{-1}	1.81×10^{-2}	6.76×10^{-2}
	3	3	9	12	7.84×10^{-2}	2.17×10^{-1}	1.35×10^{-1}	2.69×10^{-1}	1.58×10^{-2}	9.24×10^{-2}
10^{-4}	64	0	-	-	5.85×10^{-2}	3.19×10^{-1}	2.28×10^{-1}	6.93×10^{-1}	7.75×10^{-2}	4.85×10^{-1}
	6	3	-	-	3.00×10^{-2}	8.91×10^{-2}	8.03×10^{-2}	1.24×10^{-1}	4.00×10^{-3}	1.32×10^{-2}
	6	3	12	19	3.00×10^{-2}	8.10×10^{-2}	8.08×10^{-2}	1.28×10^{-1}	8.68×10^{-3}	4.49×10^{-2}
10^{-5}	82	0	-	-	5.10×10^{-2}	2.99×10^{-1}	2.02×10^{-1}	6.71×10^{-1}	6.68×10^{-2}	4.55×10^{-1}
	9	3	-	-	1.09×10^{-2}	2.75×10^{-2}	6.69×10^{-2}	9.01×10^{-2}	1.73×10^{-3}	5.74×10^{-3}
	9	3	15	26	1.27×10^{-2}	3.85×10^{-2}	6.73×10^{-2}	9.31×10^{-2}	5.38×10^{-3}	4.09×10^{-2}

Table 6.3: PPDE solution errors in the test set as a function of the retained eigenvalue energy. No debiasing, $\varepsilon = 10^{-2}, \tau = 10^{-4}, \rho(\mu) = f(\mu)$.

Again, we compute the error in the reference domain to estimate the error induced by inverting the mapping. Using hyper-reduction, the average L^2 error is 6.71×10^{-3} with a maximum of 2.76×10^{-2}. For the H^1 error, we find 1.60×10^{-2} and 4.91×10^{-2}. We conclude that the error contributions are of the same order of magnitude in this case.

Influence of the smoothing parameter

The parameter ε influences the method in two ways. Firstly, it acts as a hyperparameter that influences the fidelity and regularity of the mapping Φ. Secondly, when using $\bar{\rho} = W_2\mathrm{Bar}_\varepsilon\{\rho_i\}_{i=1}^{n_s}$, ε influences the shape of $\bar{\rho}$. Especially with no debiasing, the entropic bias (see Section 5.3.2) leads to a smoothing of $\bar{\rho}$.

When $\rho(\mu) = f(\mu)$, we know from Section 5.3.2 that $\bar{\rho} \approx \mathcal{N}(\bar{\mu} := \frac{1}{n_s}\sum_i \mu_i, \mathrm{var} + \varepsilon)$. Therefore, $\Phi_\mu^{-1}(y) \approx \sqrt{1 + \frac{\varepsilon}{\mathrm{var}}}^{-1}(y + \mu - \bar{\mu})$ and $\det D\Phi_\mu^{-1} \approx (1 + \frac{\varepsilon}{\mathrm{var}})^{-1}$. Clearly, this degrades for $\varepsilon \gg \mathrm{var}$, which has been verified numerically. Therefore, when studying the dependence of the method on ε, one has to fix the value used to compute the reference density to a value such as $\varepsilon_{\mathrm{Bar}} = 10^{-2}$. This value is not optimized extensively and values in $[\frac{1}{2}, 2] \times \varepsilon_{\mathrm{Bar}}$ are also stable.

When $\rho(\mu) \propto u(\mu)^2$, the densities are supported on the full domain, and we vary ε for all calculations. We report results in Fig. 6.10.

Firstly and importantly, we observe that the approximation quality does not strongly depend on $\varepsilon \in [10^{-3}, 10^{-2}]$. Reducing ε significantly below 10^{-3} would require all OT calculations to be moved to the log-domain, as $\min_{x,y} \exp(c(x, y)/\varepsilon)$ becomes numerically zero in double precision.

As ε becomes too large, the empirical interpolation method does not work as well, as the source terms $\{f(\mu_i)\}_i$ are no longer well aligned, leading to an increase in Q_f. As discussed in Section 6.4, the mapping can even be non-invertible in these cases. Note that the value $\varepsilon = 4 \times 10^{-2}$ corresponds to a characteristic scale of the transport problem of $\sqrt{\varepsilon} = 0.2$, which is already one fifth of the domain size.

Secondly, the method with debiasing is no longer robust: The mapping is not invertible and the error explodes. It is possible that this can be remedied by, for example, setting a minimum value for m (note that it drops to $m = 2$), but we do not explore this further. Since we rely on the entropic mappings to approximate the true transport mappings, the performance for large values of ε is not concerning.

Run times

Comparisons between the high-fidelity solver, the POD method, and the presented method depend on the size of high fidelity simulation N, the size of training- and test-set n_s and n_t, and several other factors. Depending on the choice of these parameters, one or another method might seem favorable. Regardless, we show some examples in Table 6.4. The parameters chosen in these runs are a subset of those in Table 6.1 and Table 6.3, where the corresponding errors can be found.

We clearly see the additional cost induced by the transport and registration. Hyper-reduction leads to large computational speed-up ($\approx$ 100 to 200) in the online phase by removing the dependence on N. We also see that n_m has a larger impact on the cost of the method than m. The post-processing step to map the solutions back to the physical domain is costly, as it again depends on the size of the full order model.

In several applications, this last step is not needed. Quantities of interest can be obtainable in the reference domain, so the inversion of Φ_μ^{-1} is not necessary. For example, the energy $\frac{1}{2}\int \nabla u \cdot \nabla u$ that we report here, or linear functionals of the form $l(u) = \int u f_l$ can be calculated in the reference domain with no cost depending on the dimension of the full problem. For time-dependent problems, mapping back

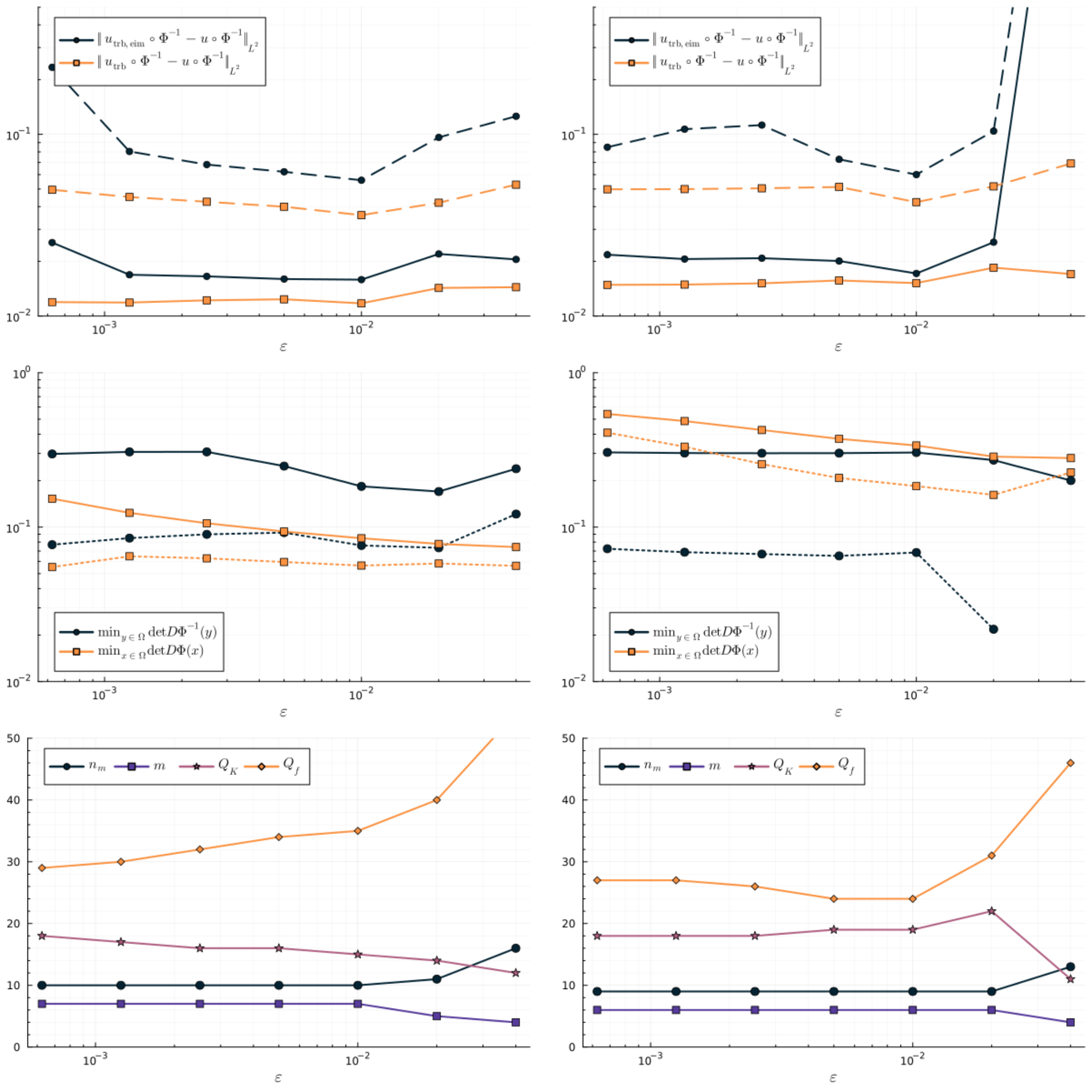

Figure 6.10: Influence of the regularization parameter ε. In all cases, $\tau = 10^{-4} = 10\tau_{\mathrm{eim}}$, $\rho(\mu) \propto u(\mu)^2$. For the left column, no debiasing is used and for the right column, debiasing is applied. First row: the relative average (solid) and maximum (dashed) L^2 error over the test set. The error in calculated in the reference domain here to separate the effect of inaccuracies when inverting Φ_μ^{-1}. Second row: minimum value of the determinant of the mapping and its inverse: average (solid) and minimum (dotted) over $\mu \in \mathcal{A}_{\mathrm{test}}$. The inverse mapping is calculated using the entropic c-transform with $\varepsilon_{\mathrm{fine}}$. Third row: number of approximation modes.

to the physical domain is only needed for diagnostics and plotting of the solution and thus usually not done at every time-step.

6.8.2 Non-linear advection equation

As a second test case, we consider the equation

$$\partial_t u(x,t) + \bar{a}(\theta) \cdot \nabla \left(u(x,t) + \gamma u(x,t)^2 \right) = \beta \Delta u(x,t), \tag{6.41}$$

	offline phase I: transport calculations (all $\mu \in \mathcal{A}_{\text{train}}$)			
	OT barycenter and potentials	boundary projection	transport modes	mapped snapshots
$\rho(\mu)$	$\bar{\rho}$ and $\{\psi^c_{\text{pre-proj.}}(\mu_i)\}_{i=1}^{n_s}$	$\{\psi^c(\mu_i)\}_{i=1}^{n_s}$	$\{\xi_j^c\}_{j=1}^{m}$	$\{u(\mu_i) \circ \Phi_{\mu_i}^{-1}\}_{i=1}^{n_s}$
$\propto u(\mu)^2$	27s	19s	33s	43s
$f(\mu)$	6.6s	19s	33s	43s

					offline phase II (all $\mu \in \mathcal{A}_{\text{train}}$)				
					Gaussian process	reduced basis	assembly	hyper-reduction	
$\rho(\mu)$	n	m	Q_K	Q_f	$\{w(\mu)\}_{i=1}^{m}$	$\{\zeta_i\}_{i=1}^{n}$ or $\{\phi_i\}_{i=1}^{n_m}$	$\int \nabla\zeta_i \cdot \nabla\zeta_j$	EIM$_K$	EIM$_f$
none	64	-	-	-	-	19s	9.4s	-	-
	82	-	-	-	-	18s	15s	-	-
$\propto u(\mu)^2$	10	7	19	24	2.3s	18s	-	86	42s
	11	10	28	30	3.6	17s	-	93s	50s
$f(\mu)$	6	3	12	19	1.3s	18s	-	83s	43s
	9	3	12	19	1.2s	18s	-	86s	43s

					online (per $\mu \in \mathcal{A}_{\text{test}}$)		post-processing (per $\mu \in \mathcal{A}_{\text{test}}$)		
					no EIM	EIM	re-mapping		
$\rho(\mu)$	n	m	Q_K	Q_f	u_{rb} or $u_{\text{trb}} \circ \Phi_\mu^{-1}$		$[\sum_{j=1}^{m} w_j(\mu)\xi_j^c]^c_{\text{pre-proj.}}$	$\Phi(\mu)$	$u_{\text{trb}}(\mu)$
none	64	-	-	-	0.14s	-	-	-	-
	82	-	-	-	0.17s	-	-	-	-
$\propto u(\mu)^2$	10	7	19	24	0.71s	4.9ms	0.24s	0.19s	0.41s
	11	10	28	30	0.95s	5.5ms	0.24s	0.20s	0.41s
$f(\mu)$	6	3	12	19	0.43s	4.5ms	0.24s	0.19s	0.40s
	9	3	15	26	0.71s	5.1ms	0.24s	0.19s	0.41s

Table 6.4: Run times for different choices of ρ. Parameters are as in Table 6.1 and Table 6.3.

where $x \in \Omega = [0,1]^2$ and $t \in [0,T]$. The advecting velocity is given by $\bar{a}(\alpha) = \frac{1}{5}(\cos\alpha, \sin\alpha)$, depending on the parameter $\alpha \in [0, 2\pi]$. The strength of the non-linearity is set to $\gamma = 10^{-2}$ and β is set to 10^{-3}. The parameter space is therefore $\mathcal{A} := [0,1] \times [0,2\pi] \ni (t,\alpha) = \mu$. As an initial condition, we choose a Gaussian centered at $(\frac{1}{2}, \frac{1}{2})$ with a variance of 5×10^{-3}. The solution is discretized using the same basis as in the previous example on a coarser 32×32 grid of quadrilaterals, using $N = 9025$ degrees of freedom. Time-integration is performed by an implicit midpoint method with time-step $\Delta t = 5 \times 10^{-2}$.

The reduced basis and transport modes are computed using the solutions at every time-step in $[0, T^{\text{train}} = \frac{4}{5}]$ for ten different values of α on a uniform grid between 0 and 2π, thus $n_s = 170$. The solutions u themselves are used as the densities $\rho(u) = u$ for the OT computations, which are performed on a 96×96 grid. We set $\varepsilon = 10^{-2}$ and use the OT barycenter of the training set as a reference density. No debiasing is used and we let $\tau_{\text{eim}} = \tau = 10^{-3}$. All other parameters are identical to the previous example.

The selected energy criterion leads to $n = 24$ for the classical RB method and $n_m = 5, m = 3$ for the proposed one.

For the hyper-reduction, the RB method with $m = 0$ requires no application of the EI method for $\bar{a}$, as the equation is already parameter-separable.

The method with transport requires EI approximations for $\det D\Phi_\mu^{-1}$, $D\Phi_\mu^{-1}\partial_t\Phi_\mu^{-1}$ $\det D\Phi_\mu^{-1}$, $D\Phi_\mu^{-1}\bar{a}(\mu)$ $\det D\Phi_\mu^{-1}$, and $K_\mu = [D\Phi_\mu^{-1}]^{-1}[D\Phi_\mu^{-1}]^{-T} \det D\Phi_\mu^{-1}$. The values of Q for these terms are $4, 4, 3$, and 4, respectively. The corresponding eigenvalue decay is shown in Fig. 6.11.

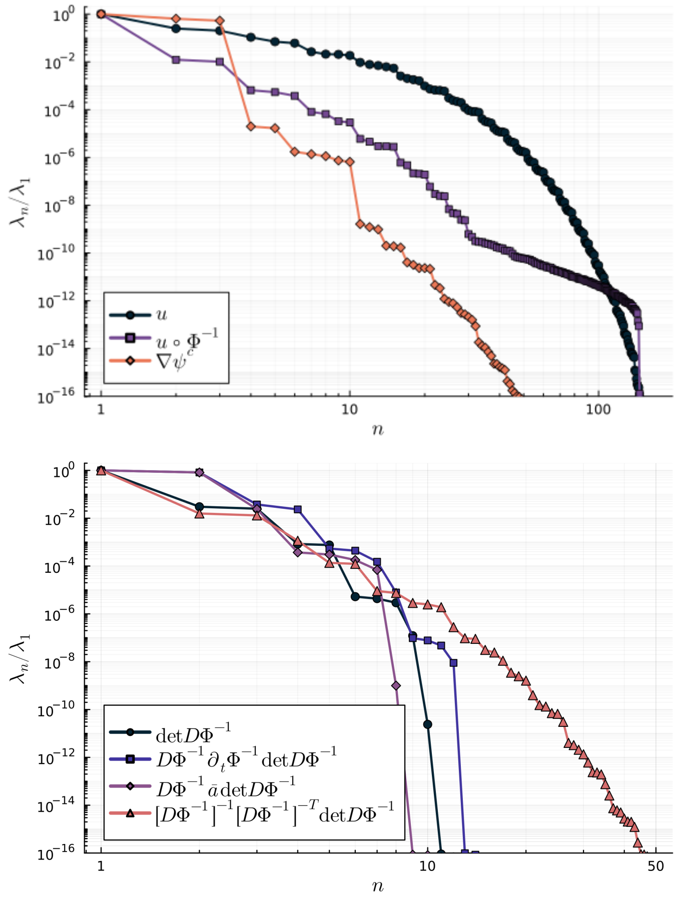

Figure 6.11: Top: Eigenvalues of the correlation matrices $\mathbb{C}^u$, $\mathbb{C}^{\Phi_* u}$, and $\mathbb{C}^{\psi}$. Bottom: Eigenvalues of the correlation matrices used in the hyper-reduction.

Figure 6.12 shows the relative L^2 errors over time for 20 values of α randomly chosen from $[0, 2\pi]$ and for $t \in [0, T^{\text{test}} = 1]$. The average approximation L^2 error of $\{u(T^{\text{test}}, \alpha_j)\}_{1 \leq j \leq 10}$ in the proposed method is 1.01×10^{-1}, while the maximum is 1.11×10^{-1}. We see that the classical RB method is only accurate for solutions close to the initial condition and for some select values of α close to those in the training set. This issue cannot be remedied by adding more reduced basis functions, since the solutions for values of α not in the training set and $t > T^{\text{train}}$ cannot be expressed by any linear combination of training snapshots. In contrast, the proposed method yields qualitatively correct results even with the low number of modes employed.

6.9 Conclusion

In the simple numerical examples we considered, the proposed method proved robust and showed the expected improvements compared to an approach without registration.

We require three central inputs from the user: An entropic regularization parameter ε, an energy cut-off criterion τ, and a choice of density ρ.

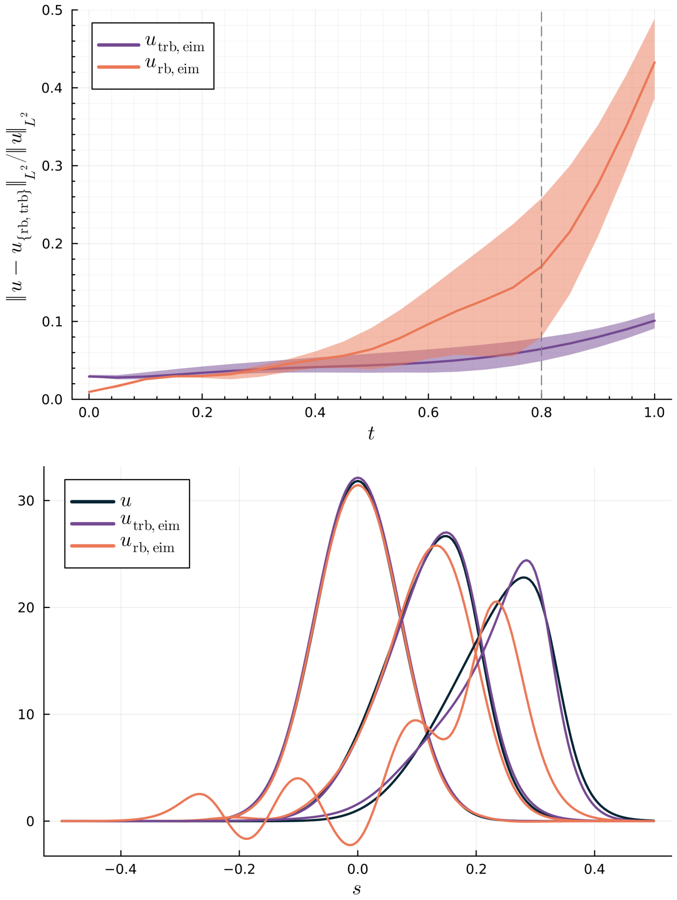

Figure 6.12: Top: Relative L^2 errors for the non-linear advection-diffusion equation as a function of time. Plotted is the average error across all ten values of α. The shaded area is bounded above and below by the maximum and minimum error. The dashed line indicates T^{train} and the beginning of the extrapolation region. Bottom: cross-sections in propagation direction (parametrized by s) for the value of α where the proposed method performs the worst and for $t \in \{0, \frac{1}{2}, 1\}$.

In order to stay close to the true OT problem, ε should be chosen small, keeping in mind that the transport plan will essentially ignore all features smaller than $\sqrt{\varepsilon}$. This parameter also has a significant influence on all transport-related computations in the offline phase, as the cost of the Sinkhorn algorithm scales with $1/\varepsilon$. The choice of τ depends on the desired accuracy versus cost of the approximation.

Selecting ρ is the most intricate issue as it has the biggest impact on regularity and fidelity of the obtained registration maps. If the PPDE solution itself is a probability density, this is a natural choice. For maximum robustness, ρ should be bounded from below on the entire domain by a positive constant. The size of this constant can have a significant effect on $\|D\Phi_\mu\|$ as discussed in Section 3.4.

6.9.1 Extensions and future work

Several possible extensions and open questions regarding the method have been mentioned throughout the text.

Smaller or no regularization

When all transported densities are strictly positive in Ω, the unregularized transport maps satisfy $\nabla\varphi(\partial\Omega) = \partial\Omega$. If Ω is also convex and the densities continuous, this guarantees $\mathcal{C}^2$ regularity of the transport maps without any entropic regularization. In this case, no post-processing of the transport maps in order to enforce the boundary conditions as in Equation (6.17) is necessary.

When employing entropic regularization, the multiscale methods from [118, 60] can drastically improve the speed of the offline calculations and allow the use of much smaller values of ε. The OT computations, in particular the application of the softmin and softmax operations (c.f. Definition 4.2), can also be performed using the same finite element spaces as used to calculate the PPDE solution itself to take advantage of higher order quadrature rules. Note also that the softmin operation provides an approximation to the c-transform (and thereby the inverse of the transport map) that is both fast and easy to implement.

We note once again that in these cases it is crucial to use choices of $\mu \mapsto \rho(\mu)$ that lead to regular transport maps in the limit of $\varepsilon \to 0$. Larger regularization parameters can provide regular maps at the cost of accuracy if this is not given, as illustrated in Figure 6.3. In many cases, the features that have to be aligned by the registration step are not supported everywhere in the domain, as seen in the examples in [73].

Representation of the transport potentials

In the present work, the potentials are represented by H^1 conforming finite element functions, which are crucially element-wise twice differentiable. We would expect the method to profit from a C^2 conforming approximation, for example using a spectral polynomial basis.

Hyper-reduction

Due to the parameter-dependence of the mapped forms, hyper-reduction methods are crucial to make the method performant in the online phase. We see in our numerical experiments that the hyper-reduction can increase the approximation error by an order of magnitude in cases where the number of reduced bases and transport modes is large and can resolve the solution well (c.f. the rightmost columns in Table 6.1 and Table 6.3). Several other hyper-reduction methods exist that have been used successfully in different applications, such as the empirical quadrature method [145]. We intend to apply these methods for comparison in future work.

6.9.2 Optimal transport and gradient mappings in model order reduction

Finally, we want to comment more generally on the strengths and weaknesses of optimal transport theory as a tool in model order reduction.

It is evident that is beneficial for a large class of reduced complexity modeling applications to move beyond linear approximation methods. A number of non-linear approaches has been proposed. These include classical methods such as wavelet bases, piece-wise polynomial approximations [50] and kernel-based methods [120]. In recent years, approaches based on neural networks have become popular in this field (for example, [113, 82, 62], to name only a few) after these methods have proved to be effective, sometimes revolutionary, in many applications ([75], to cite just one example).

The tools from optimal transport theory join this list of approaches available to tackle the challenge of slow n-width decay in reduced order problems with moving features and dominating advection effects. Several approaches have been proposed, motivated by the physically meaningful non-linear displacement interpolation [56, 73] and the information encoded in the W_2 distance between snapshots [77, 72].

What optimal transport theory can offer is an extensive theoretical framework (c.f. Chapter 3), accompanied by a number of established and optimized computational tools and algorithms (c.f. Chapter 4). Moreover, many of these results allow a physical interpretation.

However, the notion of optimality in OT will, in all generality, not coincide with what is optimal for the reduced order problem at hand. In cases where measures have non-convex support or vanishing densities, optimal transport maps can lack the regularity required for registration methods. As a result, one has to compromise by regularizing either the transport process itself in the sense of entropic regularization, or to modify the transported measures, for example to bound them away from zero.

At the same time, restricting mappings to be of gradient structure adds a number of helpful restrictions to the registration process. Note that optimization problems of the form

$$\min_{\Phi} \|u \circ \Phi^{-1} - \bar{u}\| \tag{6.42}$$

are in general non-linear and non-convex. The condition that Φ maps u into $\bar{u}$ admits a wide range of solutions. Consequently, further properties of the mapping have to be enforced with additional terms that control, for example, a norm of Φ as well as its Jacobian determinant to guarantee invertibility.

We argue that the mass transportation problem can provide a similar effect. The Jacobian determinant enters through the relation

$$\bar{\rho} = (\rho \circ \Phi)|\det D\Phi|. \tag{6.43}$$

Still this equation does not necessarily tell us much about Φ. As pointed out in [47], in the simple case where $\bar{\rho} = \mathbb{1}_{\Omega'}$ and $\rho = \mathbb{1}_{\Omega''}$ for some smooth sets Ω', Ω'', we can right compose Φ with any map S such that $S(\Omega') = \Omega'$ and $\det DS = 1$, and still satisfy Equation (6.43). If Φ transports $\bar{\rho}$ to ρ, then so does $\Phi \circ S$ and the latter map can be chosen to violate any desirable properties that Φ might have had.

Requiring Φ to an optimal transport map restricts it to the form $\Phi = \nabla\varphi$ and immediately establishes existence and uniqueness when ρ admits a density. The Monge-Ampére equation $\bar{\rho} = (\rho \circ \nabla\varphi) \det D^2\varphi$ is an elliptic problem. While non-linearity and degeneracy make its study challenging, we argue that it provides a promising starting point for registration problems. While strictly optimal transport maps can violate necessary conditions such as continuity of Φ, suitable regularization can yield almost optimal maps that avoid these singularities.

In this work, we aimed to connect the fully non-linear approximation methods based on barycentric encoding to the more classical registration methods of the form of Equation (6.42). We proposed an approach that covers the construction of the registration maps from a small number parameter choices and in a fully data-driven way. By reducing the non-linearity to the point where

$$u(\mu) \circ \Phi_\mu^{-1}(y) \approx u(\mu) \circ \left(y - \sum_j w_j \nabla \xi_j^c(y) \right) = \sum_i \tilde{u}_i(\mu) \phi_i(y) \tag{6.44}$$

is composed of two linear approximations, we obtain a method that is suited for established hyper-reduction methods and admits a performant online phase.

Appendix A

Notation

The following table repeat some of the notation and abbreviations that are repeatedly used throughout the article:

$\langle \cdot, \cdot \rangle_H$	inner product in the Hilbert space H
$.*, ./, \dots$	element-wise operations, e.g. $(\hat{a} .* \hat{b})_i = \hat{a}_i * \hat{b}_i$;
$\widehat{(\cdot)}$	collocated quantity $\hat{a}_i = a(x_i)$
$\oplus$	notation for $(\psi_\rho \oplus \psi_\sigma)(x, y) = \psi_\rho(x) + \psi_\sigma(y)$
$(\cdot)_\sharp$	push-forward operation for a density
$(\cdot)^*$	Legendre transform $f^*(y) := \sup_x (\langle x, y \rangle - f(x))$
$\mathbb{1}$	indicator function: $\mathbb{1}_{\Omega'}(x) = 1$ if $x \in \Omega'$ and zero otherwise
$\mathcal{A}$	parameter space
ρ-a.e.	ρ-almost everywhere, i.e. except on $\Omega' : \rho[\Omega'] = 0$
B	interpolation matrix used in empirical interpolation
c	cost function, usually $c(x, y) = \lvert x - y \rvert^2$
$(\cdot)^c$	c-transform, see Definition 3.4
$\mathbb{C}^u$	correlation matrix of $u_1, \dots, u_{n_s}$
$\mathrm{cdf}, \mathrm{cdf}^{[-1]}$	(inverse) cumulative distribution, see Equation (3.34)
$\mathcal{E}(n; \lambda)$	eigenvalue energy, see Eq. (6.40)
EI(M)	empirical interpolation (method), see Section 2.5
h	grid width
id, Id	identity $x \mapsto x$, identity matrix
k	Gibbs kernel $k(x, y) = \exp(-c(x, y)/\varepsilon)$
$\mathcal{L}_\mu$	parameter-dependent PDE operator, see Definition 1.1
m	number of transport modes, see Definition 6.1
$\mathcal{M}$	solution manifold, see Eq. (1.2)
$\mathrm{ME}_{\bar{\rho}}$	Monge embedding with respect to $\bar{\rho}$, see Definition 3.9
$\max^{\varepsilon}_{(\cdot)\sim u}, \min^{\varepsilon}_{(\cdot)\sim u}$	softmax and softmin operations, see Definition 4.2
n, n_m	number of reduced basis functions
n_s, n_t	number of solution snapshots in the training and test set
N	dimension of the high fidelity discretization
$\mathcal{N}(m, \mathrm{var})$	normal distribution with mean m and variance var
OT	optimal transport
$\mathcal{P}(\Omega)$	set of probability measures on Ω
POD	proper othogonal decomposition
Q	number of empirical interpolation modes

T	transport or Monge map, see Equation (3.7)
RB(M)	reduced basis (method)
u, v	elements of Hilbert space V or V_h
$\mathtt{u}, \mathtt{v}$	degree of freedom vectors $\in \mathbb{R}^N$ of $u, v \in V_h$
$\tilde{u}$	coefficients used to approximate u_{trb}, see Equation (6.7)
v	matrix eigenvector
V, V_h	Hilbert space and its discretization
w	transport mapping coefficient, see Section 6.1
$W_2(\rho, \sigma)$	optimal transport or Wasserstein distance between ρ and σ.
W_2Bar	optimal transport barycenter, see Definition 5.1
X	empirical interpolation mode, see Section 2.5
ε	entropic regularization parameter, see Eq. (4.3)
$\varepsilon_{\text{fine}}$	regularization used when inverting Φ^{-1}. Set to the order of h^2.
ζ	POD basis function
θ	empirical interpolation coefficients
κ	H^1 projection parameter, see Section 6.4. Set to $\varepsilon^{-1/2}$
λ	matrix eigenvalue, non-increasing: $\lambda_1 \geq \lambda_2 \geq \dots$
μ	parameter in $\mathcal{A}$
π	transport plan, see Definition 3.3
$\Pi(\rho, \sigma)$	admissible transport plans for ρ, σ, see Definition 3.3
ρ, σ	probability measures
$\bar{\rho}$	reference density, see Definition 3.9 and Section 6.1
Σ_m	m-unit simplex: $\{\omega_i\}_{i=1}^m \in \Sigma_m \Leftrightarrow \omega_i \geq 0 \; \forall i, \sum_i \omega_i = 1$
τ, τ_{eim}	energy criterion for POD: $1 - \mathcal{E}(n; \lambda) < \tau$ defines n
φ	$x \mapsto \frac{\lvert x\rvert^2}{2} - \psi(x)$, see Theorem 3.2
ϕ	POD basis function in the reference domain, see Equation (6.7)
Φ_μ	parameter-dependent mapping, see Equation (2.38)
ξ^c	transport mode, see Definition 6.1
Ξ	POD modes used in the EIM construction, see Section 2.5
ψ	transport potential, see Equation (3.5)
ω	barycenter weights $\{\omega_i\}_{i=1}^m \in \Sigma_m$, see Definition 5.1
Ω	domain $\subset \mathbb{R}^d$

Appendix B

Optimal transport and fluid mechanics

In this section, we want to describe two fluid dynamics applications where optimal transport problems arise. In doing so, we hope to extend the geometric picture of $\mathcal{P}(\Omega)$ introduced in Section 3.7.

B.1 Fluid dynamics as geodesic equations

It is possible to express a number of ideal fluid equations as action principles on the set of diffeomorphisms on a domain Ω, denoted by $\mathrm{Diff}(\Omega)$. This idea has been around for some time, one of the earlier references one can find is the doctoral thesis of Paul Ehrenfest [55]. In [30], this connection is attributed to Euler himself. Today, this approach is strongly linked to the works of Arnold [10]. We provide a very brief and completely formal presentation in the following. All quantities in this chapter are assumed to be smooth and we use subscript notation freely to denote dependencies on time or other parameters. Densities are assumed to be strictly positive. A thorough presentation of the variational formulation of the Euler equations in this way is given in [7, 46, 30, 29].

Let $\mathrm{Diff}(\Omega) \ni F_t : \Omega \to \Omega \subset \mathbb{R}^3$ be a time-dependent diffeomorphism at time t. In particular, F_t is the flow of a time-dependent vector field u_t:

$$\partial_t F_t(x) = u_t \circ F_t(x) \text{ where } F_0(x) = x. \tag{B.1}$$

We denote by F_t^ϵ a variation of this diffeomorphism, at is a one-parameter family of diffeomorphisms where $F_t^\epsilon\big|_{\epsilon=0} = F_t$ and $\partial_\epsilon F_t^\epsilon = v_t^\epsilon \circ F_t^\epsilon$. If we require $\partial_\epsilon \partial_t F_t^\epsilon \overset{!}{=} \partial_t \partial_\epsilon F_t^\epsilon$, we obtain the relation

$$\partial_\epsilon(u_t^\epsilon \circ F_t^\epsilon) = (\partial_\epsilon u_t^\epsilon + v_t^\epsilon \cdot \nabla u_t^\epsilon) \circ F_t^\epsilon = \partial_t(v_t^\epsilon \circ F_t^\epsilon(x)) = (\partial_t v_t^\epsilon + u_t^\epsilon \cdot \nabla v_t^\epsilon) \circ F_t^\epsilon. \tag{B.2}$$

If we define $\delta u_t := \partial_\epsilon u_t^\epsilon\big|_{\epsilon=0}$ the variation of u, then we see that variations of F_t^ϵ induce variations of u_t of the specific form

$$\delta u_t = \partial_t v_t + u_t \cdot \nabla v_t - v_t \cdot \nabla u_t = \partial_t v_t + \nabla \times (v_t \times u_t) - v_t \nabla \cdot u_t + u_t \nabla \cdot v_t, \tag{B.3}$$

where $v_t = v_t^\epsilon\big|_{\epsilon=0}$ an arbitrary, time-dependent vector field that is tangent to $\partial\Omega$ (if it was not, F_t^ϵ would no longer be a diffeomorphism on Ω). Furthermore, if we define

$\rho_t := (F_t)_\sharp \rho_0$ for some initial density ρ_0, then, by the definition of the push-forward formula for densities,

$$\begin{aligned} 0 = \partial_\epsilon \rho_0 = \partial_\epsilon((\rho_t^\epsilon \circ F_t^\epsilon) \det DF_t^\epsilon) \\ = (\partial_\epsilon \rho_t^\epsilon) \circ F_t^\epsilon + ((v_t^\epsilon \cdot \nabla \rho_t^\epsilon) + \rho_t^\epsilon \nabla \cdot v_t^\epsilon) \circ F_t^\epsilon \det DF_t^\epsilon \end{aligned} \tag{B.4}$$

by Liouville's formula as in Section 3.6. Consequently, and since $\det DF_t^\epsilon > 0$ by assumption, $\delta\rho_t := \partial_\epsilon \rho_t^\epsilon\big|_{\epsilon=0} = -\nabla \cdot (v_t \rho_t)$. By the same procedure, we obtain $\partial_t \rho_t = -\nabla \cdot (u_t \rho_t)$.

B.2 Euler equations

With these preliminary computations done, we can compute the variation of an action functional, defined in this first simple model as

$$\mathfrak{A}(t \mapsto F_t; \rho_0)_0^T := \int_0^T \int_\Omega \frac{1}{2} |u_t|^2 \mathrm{d}\rho_t . \mathrm{d}t \tag{B.5}$$

By the principle of stationary action, we set $\delta\mathfrak{A}(F_t; \rho_0) := \partial_\epsilon \mathfrak{A}(F_t^\epsilon; \rho_0)\big|_{\epsilon=0} \overset{!}{=} 0$. We obtain

$$0 = \delta\mathfrak{A}(t \mapsto F_t; \rho_0)_0^T = \int_0^T \int_\Omega \left(u \cdot \delta u \mathrm{d}\rho_t + \frac{1}{2} |u_t|^2 \mathrm{d}(\delta\rho_t) \right) \mathrm{d}t. \tag{B.6}$$

Substituting δu_t and $\delta \rho_t$, we obtain a lengthy expression under the integral that we can simplify using some basic vector calculus identities, the fact that $\delta_t v_t\big|_{t \in \{0,T\}} = 0$ (as the variation vanishes here in the stationary action principle), and the continuity equation. Ultimately, we arrive at

$$0 = \int_0^T \int_\Omega v \cdot (\partial_t u + u \cdot \nabla u) \, \mathrm{d}\rho_t \mathrm{d}t. \tag{B.7}$$

If we recall that v_t is arbitrary (and can, in particular, be localized in time), we formally obtain the pressure-less Euler equation $\partial_t u + u \cdot \nabla u = 0$.

Incompressible Euler equation Incompressible flows are generated by divergence-free vector fields, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} F_t(x) = u_t(x, F_t(x)) \text{ with } \nabla \cdot u_t = 0. \tag{B.8}$$

Consequently, $(F_t)_\sharp \rho_0 = \rho_0$. We define the set

$$\mathrm{SDiff}(\Omega; \rho_0) := \{F_t \in \mathrm{Diff} : (F_t)_\sharp \rho_0 = \rho_0\}. \tag{B.9}$$

If ρ_0 is constant, we omit it and write $\mathrm{SDiff}(\Omega)$. The condition on the push-forward adds an additional constraint to the flows F_t and their variation, which reads

$$-\nabla \cdot (u_t \rho_t) = 0 \text{ and } -\nabla \cdot (v_t \rho_t) = 0. \tag{B.10}$$

As a result, v_t is no longer fully arbitrary. While the calculation from Equation (B.6) can be done in the same way,

$$\int_0^T \int_\Omega v \cdot (\partial_t u + u \cdot \nabla u) \, \mathrm{d}\rho_t \mathrm{d}t \tag{B.11}$$

no longer allows us to conclude that $\partial_t u + u \cdot \nabla u = 0$. Instead, we can also say that $\partial_t u + u \cdot \nabla u = -\nabla p_t$, where p_t is a scalar function on Ω, since (and we have seen this calculation before in Equation (3.55))

$$\int_0^T \int_\Omega v \cdot \nabla p_t \mathrm{d}\rho_t \mathrm{d}t = 0 \quad \forall v_t : -\nabla \cdot (v_t \rho_t) = 0. \tag{B.12}$$

There is no evolution equation for the function p_t. It is naturally interpreted as the Lagrange multiplier corresponding to the incompressibility constraint.

Remark B.1. *Typically, since the density does not enter the dynamics of the problem, it is set to a constant. The incompressibility constraint reduces to* $\nabla \cdot u_t = 0$ *in this case and flows with* $\det DF_t = 1$ *are called volume-preserving.*

Before we continue, we want to repeat a positive and two negative existence results that are classical in this field.

Theorem B.1 (Existence for short times ([54], Section 15.2))**.** *For Ω compact, simply connected and either without boundary or with C^∞ boundary. For $u_0 = u_t\big|_{t=0}$ in the Sobolev space $H^s(\Omega; \mathbb{R}^d)$, $s > d/2 + 1$, divergence-free and parallel to the boundary of Ω, there exists a unique $u_t \in H^s(\Omega, \mathbb{R}^d)$ for $t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$ that solves the incompressible Euler's equations. In particular, it is a classical solution, i.e. $C^1((-\epsilon, \epsilon) \times \Omega, \mathbb{R}^3)$ and the flow F_t generated by u_t is a volume-preserving C^1 diffeomorphism.*

Theorem B.2 (Non-existence of curves with finite action for $d = 2$ ([122], Theorem 2.6))**.** *For $\Omega = [0, 1]^2$, there exists $\chi \in$* SDiff(Ω) *such that there is no curve $t \mapsto F_t$: $[0, T] \to$* SDiff(Ω) *connecting* id *to χ with $\mathfrak{A}(t \mapsto F_t)_0^T < \infty$.*

Theorem B.3 (Existence of curves with finite action for $d \geq 3$ ([121], Theorem A))**.** *For $\Omega = [0, 1]^d$, $d \geq 3$, any element in* SDiff(Ω) *can be connected with* id *by a curve $t \mapsto F_t : [0, T] \to$* SDiff$(\Omega)$ *with $\mathfrak{A}(t \mapsto F_t)_0^T < \infty$.*

Remark B.2. *As a consequence of the two preceding theorems,* $\inf \mathfrak{A}(t \mapsto F_t)_0^T$ *is not always attained if $d \geq 3$. Consider the case $d = 3$. Any element of* SDiff(Ω) *that acts as $(x, y, z) \mapsto (\chi(x, y), z)$ with χ as in Theorem B.2 cannot be connected to the identity with a curve that leaves the third component untouched, as then the action would be infinite. At the same time, however, one can reduce the action by re-scaling this non-trivial third component.*

Projection onto measure-preserving maps

This section follows [46].

One possible strategy to obtain a geodesic on SDiff$(\Omega; \rho_0)$ is the following: embed SDiff$(\Omega; \rho_0)$ in a larger space, say, $L^2(\Omega, \rho_0, \mathbb{R}^d)$. Second, since SDiff$(\Omega; \rho_0)$

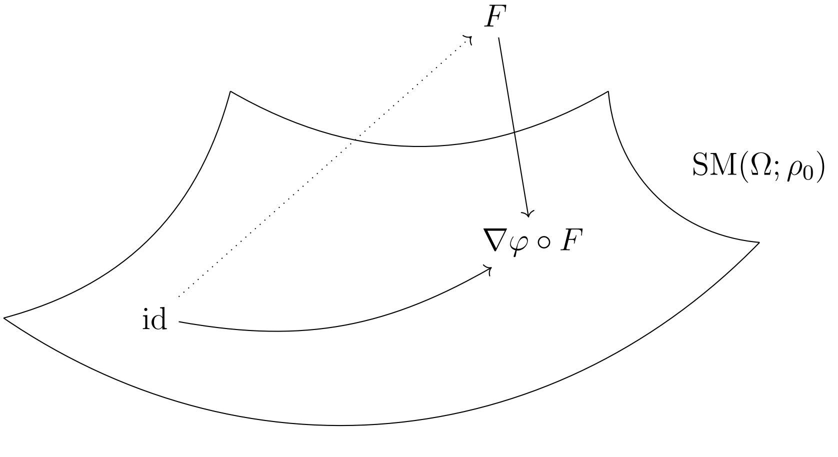

Figure B.1: Illustration of the projection of F onto $\mathrm{SM}(\Omega;\rho_0)$.

is not closed, replace it with the more workable space of *measure-preserving maps* $\mathrm{SM}(\Omega;\rho_0) := \{S \in L^2(\Omega,\rho_0,\mathbb{R}^d) : S_\sharp\rho_0 = \rho_0\}$ [31, 28]. Discretizing in time, we then follow the geodesics in this larger space for a time Δt, ignoring the constraint, and project back to $\mathrm{SDiff}(\Omega;\rho_0)$ before repeating the steps. If we take as a measure of distance (and orthogonality) the inner product of $L^2(\Omega,\rho_0,\mathbb{R}^d)$, then the optimization problem becomes the following (without changing the notation for the map F, which is now no longer assumed to be a diffeomorphism): for $F \in L^2(\Omega,\rho_0,\mathbb{R}^d)$, find

$$\min\left(\int_\Omega |S-F|^2 \mathrm{d}\rho_0 : S \in L^2(\Omega,\rho_0,\mathbb{R}^d) : S_\sharp\rho_0 = \rho_0\right) =: \mathrm{dist}^2(F,\mathrm{SM}(\Omega)). \quad \text{(B.13)}$$

The projection turns out to be an optimal transport problem with marginals $S_\sharp\rho_0 = \rho_0$ and $F_\sharp\rho_0$. As long as $F_\sharp\rho_0$ is absolutely continuous with respect to the Lebesgue measure, there exists an optimal transport map $\nabla\varphi = T_{F_\sharp\rho_0\to\rho_0}$ and $\nabla\varphi\circ F$ gives the projection of F to the measure-preserving maps, i.e. it minimizes B.13.

In [63, 95], a numerical method is constructed based on this procedure. It is a Lagrangian scheme where the density is described by particles that follow a Hamiltonian system. The distance $\mathrm{dist}^2(F,\mathrm{SM}(\Omega))$ plays the role of the potential.

B.3 Cold two-fluid plasma model

The cold two-fluid model is a simplified model to describe the dynamics of a large number of electrons and ions in a continuum approximation. We will show how one can obtain it from a physicists' derivation as we did for the Euler equations which we learned from Omar Maj. We are back to assuming a smooth setting. We begin with the action functional

$$\mathfrak{A}(t\mapsto(F_t^e,F_t^i);\rho_0,A_0)_0^T := \int_0^T\int_\Omega\left(\left(\frac{m_e}{2}|u_t^e|^2 - eA_0\cdot u_t^e\right)\mathrm{d}(F_t^e)_\sharp\rho_0 + \left(\frac{m_i}{2}|u_t^i|^2 + eA_0\cdot u_t^i\right)\mathrm{d}(F_t^i)_\sharp\rho_0\right)\mathrm{d}t. \quad \text{(B.14)}$$

Some comments are in order: F_t^e and F_t^i denote two time-dependent flows that describe the evolution of the ion and electron densities $\rho_t^e := (F_t^e)_\sharp \rho_0$ and $\rho_t^i := (F_t^i)_\sharp \rho_0$. These densities denote the number of electrons and ions per unit volume, respectively (in particular, they are not mass densities). The flows are generated by the vector fields u_t^e and u_t^i, respectively. Initially, the electron and ion densities are identical and equal to n_0. A_0 is an external vector potential, that is $\nabla \times A_0 =: B_0$ is a magnetic field. Lastly, m_e, m_i, and e are constants that denote the electron mass, ion mass, elementary charge.

The absence of a term accounting for the internal energy of the plasma gives it its name as a cold plasma model. Lastly, we assume that the ions have a charge number of one. Both of these assumptions are done in order to shorten the equations but are easy to drop.

The system comes with a constraint known as *quasi-neutrality* that couples the two flows and enforces $\rho_t^e = \rho_t^i =: \rho_t$, i.e. $(F_t^e)_\sharp \rho_0 = (F_t^i)_\sharp \rho_0$ at all times. For u_t^e and u_t^i, this constraint reads $-\nabla \cdot (e(u_t^i - u_t^e)\rho_t) = 0$, which physically means that the current density $j_t = e(u_t^i - u_t^e)\rho_t$ is divergence-free.

Carrying out the variations of $\mathfrak{A}(t \mapsto (F_t^e, F_t^i); \rho_0, A_0)_0^T$ using variations of the form

$$\delta u_t^e = \partial_t v_t^e + u_t^e \cdot \nabla v_t^e - u_t^e \cdot \nabla v_t^e \text{ and } \delta u_t^i = \partial_t v_t^i + u_t^i \cdot \nabla v_t^i - u_t^i \cdot \nabla v_t^i \tag{B.15}$$

as well as

$$\delta \rho_t^e = -\nabla \cdot (\rho_t^e v_t^e) \text{ and } \delta \rho_t^i = -\nabla \cdot (\rho_t^i v_t^i), \tag{B.16}$$

we find

$$\begin{aligned}\delta\mathfrak{A}(t \mapsto (F_t^e, F_t^i); \rho_0, A_0)_0^T \\ &= \int_0^T \int_\Omega v_t^e \cdot (-m_e \partial_t u_t^e - m_e u_t^e \cdot \nabla u_t^e - e u_t^e \times B_0)\, \mathrm{d}\rho_t^e \mathrm{d}t \\ &+ \int_0^T \int_\Omega v_t^i \cdot \left(-m_i \partial_t u_t^i - m_i u_t^i \cdot \nabla u_t^i + e u_t^i \times B_0\right) \mathrm{d}\rho_t^i \mathrm{d}t \overset{!}{=} 0. \end{aligned} \tag{B.17}$$

The calculation is rather tedious but straightforward and largely identical to the case of the Euler equations. Now, recall that v_t^e and v_t^i are not independent, but coupled through the quasi-neutrality condition that demands that variations of F_t^e and F_t^i still satisfy the condition $(F_t^{e,\epsilon})_\sharp \rho_0 = (F_t^{i,\epsilon})_\sharp \rho_0 = \rho_t$. This implies $-\nabla \cdot (e(v_t^i - v_t^e)\rho_t) = 0$. Next, note that $v_t^i = v_t^e$ is always an admissible variation. Therefore,

$$(-m_e \partial_t u_t^e - m_e u_t^e \cdot \nabla u_t^e - e u_t^e \times B_0) = -\left(-m_i \partial_t u_t^i - m_i u_t^i \cdot \nabla u_t^i + e u_t^i \times B_0\right) =: w_t \tag{B.18}$$

for some vector field w_t. Then,

$$\begin{aligned}\delta\mathfrak{A}(t \mapsto (F_t^e, F_t^i); \rho_0, A_0)_0^T = \int_0^T \int_\Omega (v_t^e - v_t^i) \cdot w_t \mathrm{d}\rho_t \mathrm{d}t = 0 \\ \forall v_t^e, v_t^i : -\nabla \cdot (e(v_t^i - v_t^e)\rho_t) = 0\end{aligned} \tag{B.19}$$

implies that $w_t = -e\nabla\phi_t$ for some scalar function ϕ_t which physically corresponds to an electric potential. Just as for the incompressible Euler equation, it plays the role of a Lagrange multiplier. Physically, it is a reasonable consequence of $\partial_t B_0 = 0$ and Faraday's law $\partial_t B_0 = -\nabla \times E$.

Projection onto quasi-neutral maps

If we want to repeat the procedure from Appendix B.2 for the two-fluid model, we require a projection that, given two maps G^i, G^e such that $G^i_\sharp\rho_0 \neq G^e_\sharp\rho_0$, returns two maps F^i, F^e where $F^i_\sharp\rho_0 = F^e_\sharp\rho_0$. A possible choice for the notion of squared distance between (G^i, G^e) and F^i, F^e is given by $\frac{1}{2}\|F^i - G^i\|^2_{L^2(\rho_0)} + \frac{1}{2}\|F^e - G^e\|^2_{L^2(\rho_0)}$, such that we arrive at the following minimization problem: Given $G^i, G^e \in L^2(\Omega, \rho_0, \mathbb{R}^d)$, find

$$\min_{F^i,F^e\in L^2(\Omega,\rho_0,\mathbb{R}^d)} \left(\int_\Omega \left(\frac{1}{2}|F^i - G^i|^2 + \frac{1}{2}|F^e - G^e|^2\right) \mathrm{d}\rho_0 : F^i_\sharp\rho_0 = F^e_\sharp\rho_0\right)$$
$$=: \mathrm{dist}^2((G^i, G^e), \mathrm{QN}(\Omega;\rho_0)). \quad \text{(B.20)}$$

Proposition B.1. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and $G^j \in L^2(\Omega, \rho_0, \mathbb{R}^d)$ for $1 \leq j \leq m$. Denote by $\{\omega_j\}_{j=1}^m$ a set of positive weights that sum to one. Assume that ρ_0 and $\{G^j_\sharp\rho_0\}_{j=1}^m$ are absolutely continuous with respect to the Lebesgue measure and their densities are bounded from above. Then,*

$$\min_{F^1,\ldots,F^m\in L^2(\Omega,\rho_0,\mathbb{R}^d)} \left(\sum_{j=1}^m \omega_j \int_\Omega |F^j - G^j|^2 \mathrm{d}\rho_0 : F^j_\sharp\rho_0 = F^1_\sharp\rho_0 \;\forall 1 \leq j \leq m\right)$$
$$= \min_{\rho\in\mathcal{P}(\Omega)} \sum_{j=1}^m \omega_j W_2(G^j_\sharp\rho_0, \rho)^2. \quad \text{(B.21)}$$

Furthermore, the minimum in Equation (B.21) *is realized by the maps $F^j = \nabla\varphi^j \circ G^j$ where $\nabla\varphi^j$ are optimal transport maps from $G^j_\sharp\rho_0$ to*

$$W_2\mathrm{Bar}(\{\omega_j, G^j_\sharp\rho_j\}_{j=1}^m) := \arg\min_{\rho\in\mathcal{P}(\Omega)} \sum_{j=1}^m \omega_j W_2(G^j_\sharp\rho_0, \rho)^2, \quad \text{(B.22)}$$

the Wasserstein barycenter of $\{G^j_\sharp\rho_0\}_j$ with weights $\{\omega_j\}_j$.

Proof. Let $G^j_\sharp\rho_0 =: \sigma_j \;\forall 1 \leq j \leq m$ and let $W_2\mathrm{Bar}(\{\omega_j, G^j_\sharp\rho_j\}_{j=1}^m) =: \bar{\rho}$. We write Equation (B.21) as

$$\min_{\substack{F^1,\ldots,F^m\in L^2(\Omega,\rho_0,\mathbb{R}^d)\\ \rho\in\mathcal{P}(\Omega)}} \left(\sum_{j=1}^m \omega_j \int_\Omega |F^j - G^j|^2 \mathrm{d}\rho_0 : F^j_\sharp\rho_0 = \rho \;\forall j\right)$$
$$= \min_{\rho\in\mathcal{P}(\Omega)} \left(\sum_j \omega_j \min_{F^j_\sharp\rho_0=\rho} \int_\Omega |F^j - G^j|^2 \mathrm{d}\rho_0\right). \quad \text{(B.23)}$$

Consider any of the inner minimization problems:

$$\min_{F^j_\sharp\rho_0=\rho} \int_\Omega |F^j - G^j|^2 \mathrm{d}\rho_0 = \min_{F^j_\sharp\rho_0=\rho} \int_{\Omega\times\Omega} |x-y|^2 \mathrm{d}((F^j, G^j)_\sharp\rho_0)(x,y) \quad \text{(B.24)}$$
$$\geq \min_{\pi\in\Pi(\sigma_j,\rho)} \int_{\Omega\times\Omega} |x-y|^2 \mathrm{d}\pi(x,y) \quad \text{(B.25)}$$
$$= W_2(\sigma_j, \rho)^2. \quad \text{(B.26)}$$

Therefore,

$$\min_{\rho\in\mathcal{P}(\Omega)}\left(\sum_j \omega_j \min_{F^j_\sharp\rho_0=\rho}\int_\Omega |F^j-G^j|^2\mathrm{d}\rho_0\right) \geq \min_{\rho\in\mathcal{P}(\Omega)}\sum_j \omega_j W_2(\sigma_j,\rho)^2 \tag{B.27}$$
$$= \sum_j \omega_j W_2(\sigma_j,\bar\rho)^2.$$

By the assumptions on σ_j for all j and the result on the regularity of the Wasserstein barycenter ([1], Theorem 5.1), the barycenter $\bar\rho$ is also absolutely continuous and bounded from above. Then, by Brenier's Theorem 3.2, the OT plan between σ_j and $\bar\rho$ is of the form $(\mathrm{id},\nabla\varphi^j)_\sharp\sigma_j$ for all j. Consequently,

$$\sum_j \omega_j W_2(\sigma_j,\bar\rho)^2 = \sum_j \omega_j \min_{\pi\in\Pi(\sigma_j,\bar\rho)}\int_{\Omega\times\Omega}|x-y|^2\mathrm{d}\pi(x,y) \tag{B.28}$$
$$= \sum_j \omega_j \int_{\Omega\times\Omega}|x-y|^2\mathrm{d}((\mathrm{id},\nabla\varphi^j)_\sharp\sigma_j)(x,y) \tag{B.29}$$
$$= \sum_j \omega_j \int_\Omega |\nabla\varphi^j\circ G^j - G^j|^2\mathrm{d}\rho_0 \tag{B.30}$$
$$\geq \min_{\rho\in\mathcal{P}(\Omega)}\left(\sum_j \omega_j \min_{F^j_\sharp\rho_0=\rho}\int_\Omega |F^j-G^j|^2\mathrm{d}\rho_0\right). \tag{B.31}$$

□

We end these considerations with another formal calculation. Consider a variation of $(F^i,F^e)^\epsilon$ with $(\delta F^i,\delta F^e)=\partial_\epsilon(F^i,F^e)^\epsilon\big|_{\epsilon=0}=(v^e,v^i)$. Recall that in this case

$$\delta\rho^e = \partial_\epsilon(F^i)^\epsilon_\sharp\rho_0\big|_{\epsilon=0} = -\nabla(v^e\rho^e) \text{ and } \delta\rho^i = -\nabla(v^i\rho^i). \tag{B.32}$$

Furthermore, for the Wasserstein barycenter $\bar\rho$ of a set of densities $\{\rho_j\}_j$, we have, from [1], Proposition 3.8,

$$\frac{1}{2}\sum_j \omega_j W_2(\rho_j,\bar\rho)^2 = \sum_j \omega_j \int \psi_j\mathrm{d}\rho_j \tag{B.33}$$

where ψ_j is the transport potential from ρ_j to $\bar\rho$, i.e. $T_{\rho_j\to\bar\rho}=\mathrm{id}-\nabla\psi_j$. Therefore, formally,

$$\partial_\epsilon \operatorname{dist}^2((F^i,F^e)^\epsilon,\mathrm{QN}(\Omega);\rho_0)\big|_{\epsilon=0} = \delta\left(\int\psi^e\mathrm{d}\rho^e+\int\psi^i\mathrm{d}\rho^i\right) \tag{B.34}$$
$$= \int\psi^e\mathrm{d}(\delta\rho^e)+\int\psi^i\mathrm{d}(\delta\rho^i) \tag{B.35}$$
$$= \int\nabla\psi^e\cdot v^e\mathrm{d}\rho^e+\int\nabla\psi^i\cdot v^i\mathrm{d}\rho^i, \tag{B.36}$$

where we used the differentiability of the W_2 distance with respect to the marginals. The potentials ψ^e and ψ^i are related by $\nabla(\psi^e)^c+\nabla(\psi^i)^c=0$. Of course, what we observe here is just a specific case of Equation (5.19):

$$\sum_j \omega_j\nabla\psi_j^c = 0.$$

In this way, the gradient functions that appear as the optimal transport potentials of the barycenter projection problem can be physically interpreted as the electric field $-e\nabla\phi$ acting on the two charged distributions with opposite sign.

Appendix C

Proof of Proposition 6.1

In this section we re-visit the one-dimensional example from Proposition 6.1, where analytical solutions can be calculated. The example is taken from appendix B of [128].

Proposition C.1. *The solutions to the equation*

$$-\partial_{xx}^2 u_\mu + \mu^2 u_\mu = 0 \tag{C.1}$$

on the domain $\Omega = (0,1)$ *with boundary conditions* $u_\mu(0) = 1$, $u_\mu(1) = 0$ *and* $\mu, \bar{\mu} \in [\mu_{\min}, \mu_{\max}] =: \mathcal{A}$, $\mu_{\max} = \epsilon^{-2}\mu_{\min}$, $\mu_{\min} > 1$, $\epsilon \in (0,1)$ *satisfy*

$$\|u_{\bar{\mu}} - u_\mu \circ T_{\rho_{\bar{\mu}} \to \rho_\mu}\|_{L^2(\Omega)} \leq \left| \frac{1}{\cosh \mu} - \frac{1}{\cosh \bar{\mu}} \right| \leq 2e^{-\mu_{\min}}, \tag{C.2}$$

where $\rho(u) = u / \int u$.

Proof. The solution manifold is given by

$$\mathcal{M} = \left\{ \frac{\cosh(\mu(1-x))}{\cosh \mu} : \mu \in \mathcal{A} \right\}. \tag{C.3}$$

In this one-dimensional example, the OT maps can be calculated analytically using cumulative density functions. Since $\rho(u) = u/\int u$, we find

$$\rho(u_\mu) =: \rho_\mu = \mu \frac{\cosh(\mu(1-x))}{\sinh \mu}. \tag{C.4}$$

Furthermore, $\operatorname{cdf}(\rho_\mu)(x) = 1 - \sinh(\mu(1-x))/\sinh\mu$, $\operatorname{cdf}(\rho_\mu)^{[-1]}(p) = 1 - \sinh^{-1}((1-p)\sinh\mu)/\mu$, and

$$T_{\rho_{\bar{\mu}} \to \rho_\mu}(y) = \operatorname{cdf}(\rho_\mu)^{[-1]} \circ \operatorname{cdf}(\rho_{\bar{\mu}})(y) = 1 - \frac{1}{\mu} \sinh^{-1}\left(\frac{\sinh \mu}{\sinh \bar{\mu}} \sinh(\bar{\mu}(1-y)) \right). \tag{C.5}$$

The map $T_{\rho_{\bar{\mu}} \to \rho_\mu}$ is a bijection as it is strictly increasing and $T_{\rho_{\bar{\mu}} \to \rho_\mu}(\partial\Omega) = \partial\Omega$. The former is a consequence of $0 < \rho_\mu < +\infty \ \forall \mu$.

Using $\cosh \circ \sinh^{-1}(x) = \sqrt{1+x^2}$, and letting $z = \bar{\mu}(1-y)$, we write

$$\|u_{\bar{\mu}} - u_\mu \circ T_{\rho_{\bar{\mu}} \to \rho_\mu}\|_{L^2(\Omega)}^2$$

$$= \frac{1}{\bar{\mu}} \int_0^{\bar{\mu}} \frac{1}{\cosh \mu^2} \left(\frac{\cosh \mu}{\cosh \bar{\mu}} \cosh z - \sqrt{1 + \frac{\sinh \mu^2}{\sinh \bar{\mu}^2} \sinh z^2} \right)^2 \mathrm{d}z \tag{C.6}$$

where $\sinh z^2$ denotes $(\sinh z)^2$. We can check using symbolic numerical software that the integrand has no local maximum, as any stationary condition at $z \neq 0$ requires either

$$\frac{\sinh \mu^2}{\sinh \bar{\mu}^2} < \frac{\cosh \mu}{\cosh \bar{\mu}} < \frac{\sinh \mu}{\sinh \bar{\mu}} \Leftrightarrow \frac{\sinh \mu}{\sinh \bar{\mu}} < \frac{\tanh \bar{\mu}}{\tanh \mu} < 1 \tag{C.7}$$

or the same relation will all inequalities reversed, either of which lead to contradiction for $\mu, \bar{\mu} > 0$. As the integrand vanishes at $z = \bar{\mu}$, its maximum value is attained at $z = 0$, i.e. $y = 1$, and we find

$$\|u_{\bar{\mu}} - u_\mu \circ T_{\rho_{\bar{\mu}} \to \rho_\mu}\|_{L^2(\Omega)} \leq |\Omega| |u_{\bar{\mu}}(1) - u_\mu \circ T_{\rho_{\bar{\mu}} \to \rho_\mu}(1)| = \left| \frac{1}{\cosh \bar{\mu}} - \frac{1}{\cosh \mu} \right|. \tag{C.8}$$

□

Remark C.1. *If we chose $\bar{\rho} = \mathrm{OTBar}\{\rho_\mu : \mu \in \mathcal{A}\}$ with uniform weights, we find $\bar{\rho} = \rho_{\bar{\mu}}$ with*

$$\bar{\mu} = \tfrac{\mu_{\max} - \mu_{\min}}{\log \mu_{\max} - \log \mu_{\min}}, \tag{C.9}$$

the logarithmic mean. The choice made in [128] is $\bar{\mu} = \sqrt{\mu_{\min}\mu_{\max}}$.

Remark C.2. *Note that*

$$T_{\rho_{\bar{\mu}} \to \rho_\mu}(y) \approx T_{\rho_{\bar{\mu}} \to \rho_\mu}(0) + y \partial_y T_{\rho_{\bar{\mu}} \to \rho_\mu}(0) = \frac{\bar{\mu}}{\mu} \frac{\tanh \mu}{\tanh \bar{\mu}} y \approx \frac{\bar{\mu}}{\mu} y \tag{C.10}$$

and this approximation is close until either $y \approx \mu/\bar{\mu}$ (when $\mu < \bar{\mu}$) or $y \approx 1$ (when $\mu > \bar{\mu}$).

The derivative of $T_{\rho_{\bar{\mu}} \to \rho_\mu}$ can take extreme values: $\partial_y T_{\rho_{\bar{\mu}} \to \rho_\mu}(1) = \frac{\bar{\mu}}{\mu} \frac{\sinh \mu}{\sinh \bar{\mu}} \approx \frac{\bar{\mu}}{\mu} e^{\mu - \bar{\mu}}$.

We now give the proof of Proposition 6.1:

Proof. We want to show that (c.f. Eq. (6.11))

$$\inf_{\xi_1^c \in \mathrm{span}\{\psi_\mu^c : \mu \in \mathcal{A}\}} \sup_{\mu \in \mathcal{A}} \inf_{\substack{w_1(\mu) : \Phi^{-1}(y) = y - w_1(\mu) \partial_y \xi_1^c(y) \\ \Phi_\mu^{-1} : \Omega \to \Omega \text{ is a bijection}}} \|u_{\bar{\mu}} - u_\mu \circ \Phi_\mu^{-1}\|_{L^2(\Omega)} \leq e^{-\mu_{\min}}(4 + \epsilon).$$

By Remark C.1, $\bar{\rho} = \rho_{\bar{\mu}}$ with $\bar{\mu} = (\mu_{\max} - \mu_{\min})/(\log \mu_{\max} - \log \mu_{\min})$. Let

$$c(\mu) = \frac{\bar{\mu} - \mu}{\mu} \text{ and } w(\mu) = -\frac{c(\mu)}{c(\mu_{\min})}. \tag{C.11}$$

We will show the bound by evaluating it at the trial function

$$\Phi_\mu^{-1}(y) := y - w(\mu) \left(T_{\rho_{\bar{\mu}} \to \rho_{\mu_{\min}}}(y) - y \right), \tag{C.12}$$

i.e. $\partial_y \xi_1^c(y) := T_{\rho_{\bar{\mu}} \to \rho_{\mu_{\min}}}(y) - y$. Note that, by the properties of the logarithmic mean, $\mu_{\min}/\epsilon \leq \bar{\mu} \leq \mu_{\min}/2\epsilon^2$. As a consequence, $-1 \leq w(\mu) \leq \epsilon$. As $T_{\rho_{\bar{\mu}} \to \rho_{\mu_{\min}}}$ is concave, Φ_μ^{-1} is concave for $\mu < \bar{\mu}$ and convex otherwise. Consequently, Φ_μ^{-1} is strictly increasing, with $\partial_y \Phi_\mu^{-1}(y) \geq \partial_y \Phi_{\mu_{\min}}^{-1}(1) > \bar{\mu}(\mu_{\min})^{-1} e^{\mu_{\min} - \bar{\mu}}$ for $\mu < \bar{\mu}$ and $\partial_y \Phi_\mu^{-1}(y) \geq \Phi_{\mu_{\max}}^{-1}(0) \geq \epsilon$ for $\mu \geq \bar{\mu}$.

Let $\delta' := \mu_{\min}/\bar{\mu}$ and write

$$\|u_{\bar{\mu}} - u_\mu \circ \Phi_\mu^{-1}\|^2_{L^2(\Omega)} \leq \int_0^{\delta'} \left(u_{\bar{\mu}}(y) - u_\mu \circ \frac{\bar{\mu}}{\mu} y\right)^2 \mathrm{d}y + \int_0^{\delta'} \left(u_\mu \circ \Phi_\mu^{-1}(y) - u_\mu \circ \frac{\bar{\mu}}{\mu} y\right)^2 \mathrm{d}y + \int_{\delta'}^1 \left(u_{\bar{\mu}}(y) - u_\mu \circ \Phi_\mu^{-1}(y)\right)^2 \mathrm{d}y. \tag{C.13}$$

For the first term, we can simplify $|u_{\bar{\mu}}(y) - u_\mu \circ (\bar{\mu}y/\mu)| = \sinh(\bar{\mu}y)|\tanh\bar{\mu} - \tanh\mu| \leq 2\sinh(\mu_{\min})|e^{-2\mu} - e^{-2\bar{\mu}}| \leq e^{-\mu_{\min}}\ \forall y \in [0, \delta']$.

For the second term, since Φ_μ^{-1} is either concave or convex, we find that

$$\left|\frac{\bar{\mu}}{\mu} y - \Phi_\mu^{-1}(y)\right| \leq \left|\frac{\bar{\mu}}{\mu} y - \partial_y \Phi_\mu^{-1}(0) y\right| \tag{C.14}$$

$$\leq \left|\frac{\bar{\mu}}{\mu} - 1 + w(\mu)\frac{\bar{\mu}}{\mu_{\min}}\frac{\tanh\mu_{\min}}{\tanh\bar{\mu}} - w(\mu)\right| \frac{\mu_{\min}}{\bar{\mu}} \tag{C.15}$$

$$= |w(\mu)| \left|(1 - \frac{\tanh\mu_{\min}}{\tanh\bar{\mu}}\right| \quad \forall y \in [0, \delta']. \tag{C.16}$$

Therefore,

$$\left|u_\mu \circ \Phi_\mu^{-1}(y) - u_\mu \circ \frac{\bar{\mu}}{\mu} y\right|_{0\leq y\leq\delta'} \leq \mu \underbrace{\max_{0\leq y\leq\delta'} \frac{\sinh(\mu(1-y))}{\cosh\mu}}_{=\tanh\mu} |w(\mu)| \left(1 - \frac{\tanh\mu_{\min}}{\tanh\bar{\mu}}\right). \tag{C.17}$$

Using the bounds on $w(\mu)$, $\bar{\mu}$, and $1 - \tanh\mu \leq 2e^{-2\mu}$, this expression is bounded by $\bar{\mu} \cdot 1 \cdot 1 \cdot 2e^{-2\mu_{\min}} = 2\bar{\mu}e^{-2\mu_{\min}}$ for $\mu \leq \bar{\mu}$ and $\mu_{\max} \cdot 1 \cdot \epsilon \cdot 2e^{-2\mu_{\min}} \leq 2\bar{\mu}e^{-2\mu_{\min}}$ otherwise.

For the third term, assume $\mu \leq \bar{\mu}$. Recall that in this case, $\Phi_\mu^{-1}(y) \leq \bar{\mu}y/\mu$ and therefore $u_\mu(\bar{\mu}y/\mu) \leq u_\mu \circ \Phi_\mu^{-1}(y)$. Since the mapping is increasing, the maximum of the integrand is reached at $y = \delta'$. We find the following chain of inequalities:

$$u_{\bar{\mu}}(\delta') = \cosh\mu_{\min} - \tanh\bar{\mu}\sinh\mu_{\min} \tag{C.18}$$

$$\leq \cosh\mu_{\min} - \tanh\mu\sinh\mu_{\min} \tag{C.19}$$

$$= u_\mu \circ \frac{\bar{\mu}}{\mu}\delta' \tag{C.20}$$

$$\leq u_\mu \circ \Phi_\mu^{-1}(\delta'). \tag{C.21}$$

Lastly, using

$$T_{\rho_{\bar{\mu}} \to \rho_{\mu_{\min}}}(\delta') = 1 - \tfrac{1}{\mu_{\min}} \sinh^{-1}\!\left(\tfrac{\sinh\mu_{\min}}{\sinh\bar{\mu}} \sinh(\bar{\mu} - \mu_{\min})\right) \geq 1 - \tfrac{1}{\mu_{\min}}, \tag{C.22}$$

we arrive, after some simplifications, at

$$\mu(1 - \Phi_\mu^{-1}(\delta')) \leq \mu - \mu_{\min} + \tfrac{\bar{\mu}-\mu}{\bar{\mu}-\mu_{\min}} \leq \mu - \mu_{\min} + 1 \tag{C.23}$$

and

$$u_\mu \circ \Phi_\mu^{-1}(\delta') \leq \cosh(\mu_{\min} - 1) - \sinh(\mu_{\min} - 1)\tanh\mu \tag{C.24}$$

$$= e^{1-\mu_{\min}} + \sinh(\mu_{\min} - 1)(1 - \tanh\mu). \tag{C.25}$$

Therefore,

$$\left|u_{\bar{\mu}}(\delta') - u_\mu \circ \Phi_\mu^{-1}(\delta')\right| \leq u_\mu \circ \Phi_\mu^{-1}(\delta') \leq e^{1-\mu_{\min}} + e^{\mu_{\min}-1}e^{-2\mu} \leq e^{-\mu_{\min}}(e+e^{-1}). \tag{C.26}$$

When $\mu \geq \bar{\mu}$, the reversed inequalities hold and we obtain

$$\left|u_{\bar{\mu}}(\delta') - u_\mu \circ \Phi_\mu^{-1}(\delta')\right| \leq u_{\bar{\mu}} \circ \Phi_{\bar{\mu}}^{-1}(\delta') = \cosh \mu_{\min} - \sinh \mu_{\min} \tanh \bar{\mu} \leq 2e^{-\mu_{\min}}. \tag{C.27}$$

Collecting all terms, we find

$$\|u_{\bar{\mu}} - u_\mu \circ \Phi_\mu^{-1}\|_{L^2(\Omega)} \leq \delta' e^{-\mu_{\min}} + \delta' 2\bar{\mu} e^{-2\mu_{\min}} + (1-\delta')(e+e^{-1})e^{-\mu_{\min}} < e^{-\mu_{\min}}(4+\epsilon), \tag{C.28}$$

using $0 < \delta' = \frac{\mu_{\min}}{\bar{\mu}} < \epsilon$ and $\mu_{\min} > 1$. □

Bibliography

[1] M. Agueh and G. Carlier, *Barycenters in the Wasserstein Space*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 904–924, https://doi.org/10.1137/100805741.

[2] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Transactions on Signal Processing, 54 (2006), pp. 4311–4322, https://doi.org/10.1109/TSP.2006.881199.

[3] A. Aldroubi, S. Li, and G. K. Rohde, *Partitioning signal classes using transport transforms for data analysis and machine learning*, Sampling Theory, Signal Processing, and Data Analysis, 19 (2021), p. 6, https://doi.org/10.1007/s43670-021-00009-z.

[4] J. Altschuler, F. Bach, A. Rudi, and J. Niles-Weed, *Massively scalable Sinkhorn distances via the Nyström method*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[5] J. Altschuler, J. Niles-Weed, and P. Rigollet, *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017.

[6] J. M. Altschuler, J. Niles-Weed, and A. J. Stromme, *Asymptotics for Semidiscrete Entropic Optimal Transport*, SIAM Journal on Mathematical Analysis, 54 (2022), pp. 1718–1741, https://doi.org/10.1137/21M1440165.

[7] L. Ambrosio and A. Figalli, *Lecture notes on variational models for incompressible Euler equations*, in Optimal Transport: Theory and Applications, C. Villani, H. Pajot, and Y. Ollivier, eds., London Mathematical Society Lecture Note Series, Cambridge University Press, Cambridge, 2014, pp. 58–71, https://doi.org/10.1017/CBO9781107297296.005.

[8] L. Ambrosio and N. Gigli, *A User's Guide to Optimal Transport*, in Modelling and Optimisation of Flows on Networks, vol. 2062, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 1–155, https://doi.org/10.1007/978-3-642-32160-3_1. Series Title: Lecture Notes in Mathematics.

[9] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein Generative Adversarial Networks*, in Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh, eds., vol. 70 of Proceedings of Machine Learning Research, PMLR, Aug. 2017, pp. 214–223.

[10] V. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Annales de l'institut Fourier, 16 (1966), pp. 319–361, `https://doi.org/10.5802/aif.233`.

[11] S. Badia and F. Verdugo, *Gridap: An extensible Finite Element toolbox in Julia*, Journal of Open Source Software, 5 (2020), p. 2520, `https://doi.org/10.21105/joss.02520`.

[12] F. Ballarin, E. Faggiano, S. Ippolito, A. Manzoni, A. Quarteroni, G. Rozza, and R. Scrofani, *Fast simulations of patient-specific haemodynamics of coronary artery bypass grafts based on a POD–Galerkin method and a vascular shape parametrization*, Journal of Computational Physics, 315 (2016), pp. 609–628, `https://doi.org/10.1016/j.jcp.2016.03.065`.

[13] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, *An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations*, Comptes Rendus Mathematique, 339 (2004), pp. 667–672, `https://doi.org/10.1016/j.crma.2004.08.006`.

[14] B. Battisti, T. Blickhan, G. Enchery, V. Ehrlacher, D. Lombardi, and O. Mula, *Wasserstein model reduction approach for parametrized flow problems in porous media*, ESAIM: Proceedings and Surveys, 73 (2023), pp. 28–47, `https://doi.org/10.1051/proc/202373028`.

[15] J.-D. Benamou, *Optimal transportation, modelling and numerical simulation*, Acta Numerica, 30 (2021), pp. 249–325, `https://doi.org/10.1017/S0962492921000040`.

[16] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numerische Mathematik, 84 (2000), pp. 375–393, `https://doi.org/10.1007/s002110050002`.

[17] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, *Iterative Bregman Projections for Regularized Transportation Problems*, Dec. 2014. arXiv:1412.5154 [math].

[18] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox, eds., *Model Reduction and Approximation: Theory and Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, July 2017, `https://doi.org/10.1137/1.9781611974829`.

[19] R. J. Berman, *Convergence Rates for Discretized Monge–Ampère Equations and Quantitative Stability of Optimal Transport*, Foundations of Computational Mathematics, 21 (2021), pp. 1099–1140, `https://doi.org/10.1007/s10208-020-09480-x`.

[20] E. Bernton, P. Ghosal, and M. Nutz, *Entropic Optimal Transport: Geometry and Large Deviations*, Jan. 2022. arXiv:2102.04397 [math].

[21] D. P. Bertsekas, *The auction algorithm: A distributed relaxation method for the assignment problem*, Annals of Operations Research, 14 (1988), pp. 105–123, https://doi.org/10.1007/BF02186476.

[22] D. P. Bertsekas, *Auction Algorithms*, in Encyclopedia of Optimization, C. A. Floudas and P. M. Pardalos, eds., Springer US, Boston, MA, 2008, pp. 128–132, https://doi.org/10.1007/978-0-387-74759-0_22.

[23] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Convergence Rates for Greedy Algorithms in Reduced Basis Methods*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 1457–1472, https://doi.org/10.1137/100795772.

[24] J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford, *Towards Optimal Running Times for Optimal Transport*, Jan. 2020. arXiv:1810.07717 [cs].

[25] T. Blickhan, *A registration method for reduced basis problems using linear optimal transport*, Sept. 2023. arXiv:2304.14884 [cs, math].

[26] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, vol. 44 of Springer Series in Computational Mathematics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, https://doi.org/10.1007/978-3-642-36519-5.

[27] N. Bonneel, G. Peyré, and M. Cuturi, *Wasserstein barycentric coordinates: histogram regression using optimal transport*, ACM Transactions on Graphics, 35 (2016), pp. 1–10, https://doi.org/10.1145/2897824.2925918.

[28] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on Pure and Applied Mathematics, 44 (1991), pp. 375–417, https://doi.org/10.1002/cpa.3160440402.

[29] Y. Brenier, *Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations*, Communications on Pure and Applied Mathematics, 52 (1999), pp. 411–452, https://doi.org/10.1002/(SICI)1097-0312(199904)52:4<411::AID-CPA1>3.0.CO;2-3.

[30] Y. Brenier, *Some Variational and Stochastic Methods for the Euler Equations of Incompressible Fluid Dynamics and Related Models*, in Stochastic Geometric Mechanics, S. Albeverio, A. B. Cruzeiro, and D. Holm, eds., vol. 202, Springer International Publishing, Cham, 2017, pp. 169–189, https://doi.org/10.1007/978-3-319-63453-1_8. Series Title: Springer Proceedings in Mathematics & Statistics.

[31] Y. Brenier and W. Gangbo, *\$L^p\$ Approximation of maps by diffeomorphisms*, Calculus of Variations and Partial Differential Equations, 16 (2003), pp. 147–164, https://doi.org/10.1007/s005260100144.

[32] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici, *A priori convergence of the Greedy algorithm for the parametrized reduced basis method*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 595–603, https://doi.org/10.1051/m2an/2011056.

[33] T. Bui and N. V. Phong, *A Short Note on RAS Method*, Advances in Management and Applied Economics, 3 (2013). https://EconPapers.repec.org/RePEc:spt:admaec:v:3:y:2013:i:4:f:3_4_12.

[34] L. A. Caffarelli, *The regularity of mappings with a convex potential*, Journal of the American Mathematical Society, 5 (1992), pp. 99–104, https://doi.org/10.1090/S0894-0347-1992-1124980-8.

[35] N. Cagniart, Y. Maday, and B. Stamm, *Model Order Reduction for Problems with Large Convection Effects*, in Contributions to Partial Differential Equations and Applications, B. N. Chetverushkin, W. Fitzgibbon, Y. Kuznetsov, P. Neittaanmäki, J. Periaux, and O. Pironneau, eds., vol. 47, Springer International Publishing, Cham, 2019, pp. 131–150, https://doi.org/10.1007/978-3-319-78325-3_10. Series Title: Computational Methods in Applied Sciences.

[36] E. Calabi, *Improper affine hyperspheres of convex type and a generalization of a theorem by K. Jörgens.*, Michigan Mathematical Journal, 5 (1958), https://doi.org/10.1307/mmj/1028998055.

[37] K. Carlberg, C. Bou-Mosleh, and C. Farhat, *Efficient non-linear model reduction via a least-squares Petrov–Galerkin projection and compressive tensor approximations*, International Journal for Numerical Methods in Engineering, 86 (2011), pp. 155–181, https://doi.org/10.1002/nme.3050.

[38] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer, *Convergence of Entropic Schemes for Optimal Transport and Gradient Flows*, SIAM Journal on Mathematical Analysis, 49 (2017), pp. 1385–1418, https://doi.org/10.1137/15M1050264.

[39] S. Chaturantabut and D. C. Sorensen, *Nonlinear Model Reduction via Discrete Empirical Interpolation*, SIAM Journal on Scientific Computing, 32 (2010), pp. 2737–2764, https://doi.org/10.1137/090766498.

[40] K. Cheng, S. Aeron, M. C. Hughes, and E. L. Miller, *Dynamical Wasserstein Barycenters for Time-series Modeling*, in Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, eds., vol. 34, Curran Associates, Inc., 2021, pp. 27991–28003.

[41] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *Scaling algorithms for unbalanced optimal transport problems*, Mathematics of Computation, 87 (2018), pp. 2563–2609, https://doi.org/10.1090/mcom/3303.

[42] P. Colombo, G. Staerman, C. Clavel, and P. Piantanida, *Automatic Text Evaluation through the Lens of Wasserstein Barycenters*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Nov. 2021, Association for Computational Linguistics, pp. 10450–10466, `https://doi.org/10.18653/v1/2021.emnlp-main.817`.

[43] J. B. Conway, *A Course in Functional Analysis*, vol. 96 of Graduate Texts in Mathematics, Springer New York, New York, NY, 2007, `https://doi.org/10.1007/978-1-4757-4383-8`.

[44] M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*, in Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., 2013.

[45] M. Cuturi and A. Doucet, *Fast Computation of Wasserstein Barycenters*, in Proceedings of the 31st International Conference on Machine Learning, vol. 32 of Proceedings of Machine Learning Research, Beijing, China, 2014, PMLR, pp. 685–693.

[46] S. Daneri and A. Figalli, *Variational models for the incompressible Euler equations*, in HCDTE Lecture Notes. Part II. Nonlinear HYperboliC PDEs, Dispersive and Transport Equations, vol. 7 of AIMS on Applied Mathematics, American Institute of Mathematical Sciences, Springfield, 2013, pp. 1–50.

[47] G. De Philippis, *Regularity of Optimal Transport Maps and Applications*, Scuola Normale Superiore, Pisa, 2013, `https://doi.org/10.1007/978-88-7642-458-8`.

[48] G. De Philippis and A. Figalli, *The Monge–Ampère equation and its link to optimal transportation*, Bulletin of the American Mathematical Society, 51 (2014), pp. 527–580, `https://doi.org/10.1090/S0273-0979-2014-01459-4`.

[49] G. De Philippis and A. Figalli, *Partial regularity for optimal transport maps*, Publications mathématiques de l'IHÉS, 121 (2015), pp. 81–112, `https://doi.org/10.1007/s10240-014-0064-7`.

[50] R. A. DeVore, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150, `https://doi.org/10.1017/S0962492900002816`.

[51] M.-H. Do, J. Feydy, and O. Mula, *Approximation and Structured Prediction with Sparse Wasserstein Barycenters*, Feb. 2023. arXiv:2302.05356 [cs, math].

[52] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, *Euclidean Distance Matrices: Essential theory, algorithms, and applications*, IEEE Signal Processing Magazine, 32 (2015), pp. 12–30, `https://doi.org/10.1109/MSP.2015.2398954`.

[53] M. Drohmann, B. Haasdonk, and M. Ohlberger, *Reduced Basis Approximation for Nonlinear Parametrized Evolution Equations based on Empirical Operator Interpolation*, SIAM Journal on Scientific Computing, 34 (2012), pp. A937–A969, https://doi.org/10.1137/10081157X.

[54] D. G. Ebin and J. Marsden, *Groups of Diffeomorphisms and the Motion of an Incompressible Fluid*, The Annals of Mathematics, 92 (1970), p. 102, https://doi.org/10.2307/1970699.

[55] P. Ehrenfest, *Die Bewegung Starrer Körper in Flüssigkeiten und die Mechanik von Hertz*, PhD thesis, Technische Universität Wien, Wien, 1904.

[56] V. Ehrlacher, D. Lombardi, O. Mula, and F.-X. Vialard, *Nonlinear model reduction on metric spaces. Application to one-dimensional conservative PDEs in Wasserstein spaces*, Feb. 2020. arXiv:1909.06626 [cs, math].

[57] L. C. Evans, *Partial Differential Equations and Monge-Kantorovich Mass Transfer*, Current Developments in Mathematics, 1997 (1997), pp. 65–126, https://doi.org/10.4310/CDM.1997.v1997.n1.a2.

[58] R. Everson and L. Sirovich, *Karhunen–Loève procedure for gappy data*, Journal of the Optical Society of America A, 12 (1995), pp. 1657–1664, https://doi.org/10.1364/JOSAA.12.001657. Publisher: Optica Publishing Group.

[59] L. Fedeli, A. Huebl, F. Boillod-Cerneux, T. Clark, K. Gott, C. Hillairet, S. Jaure, A. Leblanc, R. Lehe, A. Myers, C. Piechurski, M. Sato, N. Zaim, W. Zhang, J.-L. Vay, and H. Vincenti, *Pushing the Frontier in the Design of Laser-Based Electron Accelerators with Groundbreaking Mesh-Refined Particle-In-Cell Simulations on Exascale-Class Supercomputers*, in SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, Nov. 2022, IEEE, pp. 1–12, https://doi.org/10.1109/SC41404.2022.00008.

[60] J. Feydy, *Geometric data analysis, beyond convolutions*, PhD thesis, Université Paris-Saclay, 2020.

[61] A. Figalli, *Regularity Properties of Optimal Maps Between Nonconvex Domains in the Plane*, Communications in Partial Differential Equations, 35 (2010), pp. 465–479, https://doi.org/10.1080/03605300903307673.

[62] S. Fresca, L. Dede', and A. Manzoni, *A Comprehensive Deep Learning-Based Approach to Reduced Order Modeling of Nonlinear Time-Dependent Parametrized PDEs*, Journal of Scientific Computing, 87 (2021), p. 61, https://doi.org/10.1007/s10915-021-01462-7.

[63] T. O. Gallouët and Q. Mérigot, *A Lagrangian Scheme à la Brenier for the Incompressible Euler Equations*, Foundations of Computational Mathematics, 18 (2018), pp. 835–865, https://doi.org/10.1007/s10208-017-9355-y.

[64] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, *Sample Complexity of Sinkhorn Divergences*, in Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, vol. 89 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 1574–1583.

[65] E. Geoffroy Saint-Hilaire, *Lettres écrites d'Egypte à Cuvier*, vol. 1, Librairie Hachette, Paris, 1901.

[66] N. Gigli, *On Hölder continuity-in-time of the optimal transport map towards measures along a curve*, Proceedings of the Edinburgh Mathematical Society, 54 (2011), pp. 401–409, `https://doi.org/10.1017/S001309150800117X`.

[67] M. Goldman and F. Otto, *A variational proof of partial regularity for optimal transportation maps*, Annales scientifiques de l'École normale supérieure, 53 (2020), pp. 1209–1233, `https://doi.org/10.24033/asens.2444`.

[68] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, in Advances in Neural Information Processing Systems, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds., vol. 27, Curran Associates, Inc., 2014.

[69] C. Greif and K. Urban, *Decay of the Kolmogorov N -width for wave problems*, Applied Mathematics Letters, 96 (2019), pp. 216–222, `https://doi.org/10.1016/j.aml.2019.05.013`.

[70] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, `https://doi.org/10.1007/978-3-642-56468-0`.

[71] R. Hug, E. Maitre, and N. Papadakis, *On the convergence of augmented Lagrangian method for optimal transport between nonnegative densities*, Journal of Mathematical Analysis and Applications, 485 (2020), p. 123811, `https://doi.org/10.1016/j.jmaa.2019.123811`.

[72] A. Iollo and D. Lombardi, *Advection modes by optimal mass transfer*, Physical Review E, 89 (2014), p. 022923, `https://doi.org/10.1103/PhysRevE.89.022923`.

[73] A. Iollo and T. Taddei, *Mapping of coherent structures in parameterized flows by learning optimal transportation with Gaussian models*, Journal of Computational Physics, 471 (2022), p. 111671, `https://doi.org/10.1016/j.jcp.2022.111671`.

[74] H. Janati, M. Cuturi, and A. Gramfort, *Debiased Sinkhorn barycenters*, in Proceedings of the 37th International Conference on Machine Learning, vol. 119 of PMLR, 2020.

[75] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J.

Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, *Highly accurate protein structure prediction with AlphaFold*, Nature, 596 (2021), pp. 583–589, https://doi.org/10.1038/s41586-021-03819-2.

[76] L. V. Kantorovich, *On a Problem of Monge*, Journal of Mathematical Sciences, 133 (2006), pp. 1383–1383, https://doi.org/10.1007/s10958-006-0050-9.

[77] M. Khamlich, F. Pichi, and G. Rozza, *Optimal Transport-inspired Deep Learning Framework for Slow-Decaying Problems: Exploiting Sinkhorn Loss and Wasserstein Kernel*, Aug. 2023. arXiv:2308.13840 [cs, math].

[78] J. Kitagawa, Q. Mérigot, and B. Thibert, *Convergence of a Newton algorithm for semi-discrete optimal transport*, Journal of the European Mathematical Society, 21 (2019), pp. 2603–2651, https://doi.org/10.4171/JEMS/889.

[79] J. Kosowsky and A. Yuille, *The invisible hand algorithm: Solving the assignment problem with statistical physics*, Neural Networks, 7 (1994), pp. 477–490, https://doi.org/10.1016/0893-6080(94)90081-7.

[80] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, *From Word Embeddings To Document Distances*, in Proceedings of the 32nd International Conference on Machine Learning, F. Bach and D. Blei, eds., vol. 37 of Proceedings of Machine Learning Research, Lille, France, July 2015, PMLR, pp. 957–966.

[81] H. Lavenant, *Unconditional convergence for discretizations of dynamical optimal transport*, Mathematics of Computation, 90 (2020), pp. 739–786, https://doi.org/10.1090/mcom/3567.

[82] K. Lee and K. T. Carlberg, *Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders*, Journal of Computational Physics, 404 (2020), p. 108973, https://doi.org/10.1016/j.jcp.2019.108973.

[83] T. Lin, N. Ho, and M. Jordan, *On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., vol. 97 of Proceedings of Machine Learning Research, PMLR, June 2019, pp. 3982–3991.

[84] F. Léger, *A Gradient Descent Perspective on Sinkhorn*, Applied Mathematics & Optimization, 84 (2021), pp. 1843–1855, https://doi.org/10.1007/s00245-020-09697-w.

[85] C. Léonard, *From the Schrödinger problem to the Monge–Kantorovich problem*, Journal of Functional Analysis, 262 (2012), pp. 1879–1920, https://doi.org/10.1016/j.jfa.2011.11.026.

[86] C. Léonard and ,Modal-X. Université Paris Ouest, Bât. G, 200 av. de la République. 92001 Nanterre, *A survey of the Schrödinger problem and some of its connections with optimal transport*, Discrete & Continuous Dynamical Systems - A, 34 (2014), pp. 1533–1574, https://doi.org/10.3934/dcds.2014.34.1533.

[87] B. Lévy, *A Numerical Algorithm for L $_2$ Semi-Discrete Optimal Transport in 3D*, ESAIM: Mathematical Modelling and Numerical Analysis, 49 (2015), pp. 1693–1715, https://doi.org/10.1051/m2an/2015055.

[88] B. Lévy and E. L. Schwindt, *Notions of optimal transport theory and how to implement them on a computer*, Computers & Graphics, 72 (2018), pp. 135–148, https://doi.org/10.1016/j.cag.2018.01.009.

[89] S. A. Molchanov, *Diffusion processes and Riemannian geometry*, Russian Mathematical Surveys, 30 (1975), pp. 1–63, https://doi.org/10.1070/RM1975v030n01ABEH001400.

[90] P. Monk, *Finite Element Methods for Maxwell's Equations*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, UK, 2003.

[91] C. Moosmüller and A. Cloninger, *Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations*, Information and Inference: A Journal of the IMA, 12 (2023), pp. 363–389, https://doi.org/10.1093/imaiai/iaac023.

[92] M. Mueller, S. Aeron, J. M. Murphy, and A. Tasissa, *Geometric Sparse Coding in Wasserstein Space*, Oct. 2022. arXiv:2210.12135 [cs, eess, math, stat].

[93] Q. Mérigot, *A Multiscale Approach to Optimal Transport*, Computer Graphics Forum, 30 (2011), pp. 1583–1592, https://doi.org/10.1111/j.1467-8659.2011.02032.x.

[94] Q. Mérigot, A. Delalande, and F. Chazal, *Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space*, in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, vol. 108 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 3186–3196.

[95] Q. Mérigot and J.-M. Mirebeau, *Minimal geodesics along volume preserving maps, through semi-discrete optimal transport*, May 2015. arXiv:1505.03306 [math].

[96] G. Nader and G. Guennebaud, *Instant transport maps on 2D grids*, ACM Transactions on Graphics, 37 (2018), pp. 1–13, https://doi.org/10.1145/3272127.3275091.

[97] N. J. Nair and M. Balajewicz, *Transported snapshot model order reduction approach for parametric, steady-state fluid flows containing parameter-dependent shocks*, International Journal for Numerical Methods in Engineering, 117 (2019), pp. 1234–1262, `https://doi.org/10.1002/nme.5998`.

[98] A. Natale and G. Todeschi, *A Mixed Finite Element Discretization of Dynamical Optimal Transport*, Journal of Scientific Computing, 91 (2022), p. 38, `https://doi.org/10.1007/s10915-022-01821-y`.

[99] S. Nowozin, *Streaming Log-sum-exp Computation*, May 2016. https://www.nowozin.net/sebastian/blog/streaming-log-sum-exp-computation.html.

[100] M. Nutz and J. Wiesel, *Entropic Optimal Transport: Convergence of Potentials*, Oct. 2021. arXiv:2104.11720 [math].

[101] M. Ohlberger and S. Rave, *Reduced Basis Methods: Success, Limitations and Future Challenges*, in Proceedings of the Conference Algoritmy, 2016, pp. 1–12.

[102] F. Otto, *The geometry of dissipative evolution equations: the porous medium equation*, Communications in Partial Differential Equations, 26 (2001), pp. 101–174, `https://doi.org/10.1081/PDE-100002243`.

[103] V. M. Panaretos and Y. Zemel, *An Invitation to Statistics in Wasserstein Space*, SpringerBriefs in Probability and Mathematical Statistics, Springer International Publishing, Cham, 2020, `https://doi.org/10.1007/978-3-030-38438-8`.

[104] N. Papadakis, G. Peyré, and E. Oudet, *Optimal Transport with Proximal Splitting*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 212–238, `https://doi.org/10.1137/130920058`.

[105] B. Peherstorfer, Z. Drmač, and S. Gugercin, *Stability of Discrete Empirical Interpolation and Gappy Proper Orthogonal Decomposition with Randomized and Deterministic Sampling Points*, SIAM Journal on Scientific Computing, 42 (2020), pp. A2837–A2864, `https://doi.org/10.1137/19M1307391`.

[106] G. Peyré and M. Cuturi, *Computational Optimal Transport*, vol. 11:5-6 of Foundations and Trends in Machine Learning, now Publishers, Delft, The Netherlands, 2019.

[107] A.-A. Pooladian and J. Niles-Weed, *Entropic estimation of optimal transport maps*, May 2022. arXiv:2109.12004 [math, stat].

[108] A. Quarteroni, A. Manzoni, and F. Negri, *Reduced Basis Methods for Partial Differential Equations*, vol. 92 of UNITEXT, Springer International Publishing, Cham, 2016, `https://doi.org/10.1007/978-3-319-15431-2`.

[109] A. Ramdas, N. Trillos, and M. Cuturi, *On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests*, Entropy, 19 (2017), p. 47, https://doi.org/10.3390/e19020047.

[110] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning, MIT Press, Cambridge, Mass, 2006. OCLC: ocm61285753.

[111] P. Rigollet and J. Weed, *Entropic optimal transport is maximum-likelihood deconvolution*, Comptes Rendus Mathematique, 356 (2018), pp. 1228–1235, https://doi.org/10.1016/j.crma.2018.10.010.

[112] D. Rim, B. Peherstorfer, and K. T. Mandli, *Manifold Approximations via Transported Subspaces: Model Reduction for Transport-Dominated Problems*, SIAM Journal on Scientific Computing, 45 (2023), pp. A170–A199, https://doi.org/10.1137/20M1316998.

[113] F. Romor, G. Stabile, and G. Rozza, *Non-linear Manifold Reduced-Order Models with Convolutional Autoencoders and Reduced Over-Collocation Method*, Journal of Scientific Computing, 94 (2023), p. 74, https://doi.org/10.1007/s10915-023-02128-2.

[114] G. Rozza, D. B. P. Huynh, and A. T. Patera, *Reduced Basis Approximation and a Posteriori Error Estimation for Affinely Parametrized Elliptic Coercive Partial Differential Equations: Application to Transport and Continuum Mechanics*, Archives of Computational Methods in Engineering, 15 (2008), pp. 229–275, https://doi.org/10.1007/s11831-008-9019-9.

[115] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, vol. 87 of Progress in Nonlinear Differential Equations and Their Applications, Springer International Publishing, Cham, 2015, https://doi.org/10.1007/978-3-319-20828-2.

[116] F. Santambrogio, {*Euclidean, metric, and Wasserstein*} *gradient flows: an overview*, Bulletin of Mathematical Sciences, 7 (2017), pp. 87–154, https://doi.org/10.1007/s13373-017-0101-1.

[117] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngolè, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck, *Wasserstein Dictionary Learning: Optimal Transport-Based Unsupervised Nonlinear Dictionary Learning*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 643–678, https://doi.org/10.1137/17M1140431.

[118] B. Schmitzer, *A Sparse Multiscale Algorithm for Dense Optimal Transport*, Journal of Mathematical Imaging and Vision, 56 (2016), pp. 238–259, https://doi.org/10.1007/s10851-016-0653-9.

[119] B. Schmitzer, *Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems*, SIAM Journal on Scientific Computing, 41 (2019), pp. A1443–A1481, https://doi.org/10.1137/16M1106018. _eprint: https://doi.org/10.1137/16M1106018.

[120] B. Schölkopf, A. Smola, and K.-R. Müller, *Kernel principal component analysis*, in Artificial Neural Networks — ICANN'97, G. Goos, J. Hartmanis, J. Van Leeuwen, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, eds., vol. 1327, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997, pp. 583–588, `https://doi.org/10.1007/BFb0020217`. Series Title: Lecture Notes in Computer Science.

[121] A. I. Shnirelman, *Attainable diffeomorphisms*, Geometric and Functional Analysis, 3 (1993), pp. 279–294, `https://doi.org/10.1007/BF01895690`.

[122] A. I. Shnirelman, *Generalized fluid flows, their approximation and applications*, Geometric and Functional Analysis, 4 (1994), pp. 586–620, `https://doi.org/10.1007/BF01896409`.

[123] B. Simon, *Advanced Complex Analysis*, American Mathematical Society, Providence, Rhode Island, 2015, `https://doi.org/10.1090/simon/002.2`.

[124] S. P. Singh, A. Hug, A. Dieuleveut, and M. Jaggi, *Context Mover's Distance & Barycenters: Optimal Transport of Contexts for Building Representations*, in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, S. Chiappa and R. Calandra, eds., vol. 108 of Proceedings of Machine Learning Research, PMLR, Aug. 2020, pp. 3437–3449.

[125] R. Sinkhorn, *A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices*, The Annals of Mathematical Statistics, 35 (1964), pp. 876–879, `https://doi.org/10.1214/aoms/1177703591`.

[126] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, *Convolutional wasserstein distances: efficient optimal transportation on geometric domains*, ACM Transactions on Graphics, 34 (2015), pp. 1–11, `https://doi.org/10.1145/2766963`.

[127] T. J. Sullivan, *Bayesian inverse problems in function spaces.* IGDK 1754 Compact Course Lecture notes, Sept. 2019.

[128] T. Taddei, *A Registration Method for Model Order Reduction: Data Compression and Geometry Reduction*, SIAM Journal on Scientific Computing, 42 (2020), pp. A997–A1027, `https://doi.org/10.1137/19M1271270`.

[129] T. Taddei, *An optimization-based registration approach to geometry reduction*, Nov. 2022. arXiv:2211.10275 [cs, math].

[130] T. Taddei and L. Zhang, *A discretize-then-map approach for the treatment of parameterized geometries in model order reduction*, Computer Methods in Applied Mechanics and Engineering, 384 (2021), p. 113956, `https://doi.org/10.1016/j.cma.2021.113956`.

[131] T. Taddei and L. Zhang, *Space-time registration-based model reduction of parameterized one-dimensional hyperbolic PDEs*, ESAIM: Mathematical Modelling and Numerical Analysis, 55 (2021), pp. 99–130, `https://doi.org/10.1051/m2an/2020073`.

[132] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, *Overrelaxed Sinkhorn–Knopp Algorithm for Regularized Optimal Transport*, Algorithms, 14 (2021), p. 143, https://doi.org/10.3390/a14050143.

[133] S. Torregrosa, V. Champaney, A. Ammar, V. Herbert, and F. Chinesta, *Surrogate parametric metamodel based on Optimal Transport*, Mathematics and Computers in Simulation, 194 (2022), pp. 36–63, https://doi.org/10.1016/j.matcom.2021.11.010.

[134] A. Trouve, *Diffeomorphisms Groups and Pattern Matching in Image Analysis*, International Journal of Computer Vision, 28 (1998), pp. 213–221, https://doi.org/https://doi.org/10.1023/A:1008001603737.

[135] J. Urbas, *Mass Transfer Problems*, Tech. Report 41, Institut für angewandte Mathematik der Universität Bonn, 1998.

[136] F. Verdugo and S. Badia, *The software design of Gridap: A Finite Element package based on the Julia JIT compiler*, Computer Physics Communications, 276 (2022), p. 108341, https://doi.org/10.1016/j.cpc.2022.108341.

[137] A. M. Vershik, *Long History of the Monge-Kantorovich Transportation Problem: (Marking the centennial of L.V. Kantorovich's birth!)*, The Mathematical Intelligencer, 35 (2013), pp. 1–9, https://doi.org/10.1007/s00283-013-9380-x.

[138] F.-X. Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*. May 2019.

[139] C. Villani, *Optimal transport: old and new*, no. 338 in Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 2009.

[140] C. Villani, *Topics in optimal transportation*, no. 58 in Graduate studies in mathematics, American Mathematical Society, Providence, Rhode Island, reprinted with corrections ed., 2016.

[141] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, *A Linear Optimal Transportation Framework for Quantifying and Visualizing Variations in Sets of Images*, International Journal of Computer Vision, 101 (2013), pp. 254–269, https://doi.org/10.1007/s11263-012-0566-z.

[142] G. Welper, *Interpolation of Functions with Parameter Dependent Jumps by Transformed Snapshots*, SIAM Journal on Scientific Computing, 39 (2017), pp. A1225–A1250, https://doi.org/10.1137/16M1059904.

[143] M. Werenski, R. Jiang, A. Tasissa, S. Aeron, and J. M. Murphy, *Measure Estimation in the Barycentric Coding Model*, in Proceedings of the 39th International Conference on Machine Learning, vol. 162 of Proceedings of Machine Learning Research, Baltimore, Maryland, USA, 2022, PMLR, pp. 23781–23803.

[144] H. Xu, W. Wang, W. Liu, and L. Carin, *Distilled Wasserstein Learning for Word Embedding and Topic Modeling*, in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., vol. 31, Curran Associates, Inc., 2018.

[145] M. Yano and A. T. Patera, *An LP empirical quadrature procedure for reduced basis treatment of parametrized nonlinear PDEs*, Computer Methods in Applied Mechanics and Engineering, 344 (2019), pp. 1104–1123, `https://doi.org/10.1016/j.cma.2018.02.028`.