

Mathematical Foundations of Interacting Multi-Particle Systems for Optimization

Konstantin Riedl

Technical University of Munich
TUM School of Computation, Information and Technology
Department of Mathematics
Chair of Applied Numerical Analysis
Munich Center for Machine Learning
Institute for Ethics in Artificial Intelligence



Technical University of Munich
TUM School of Computation, Information and Technology

Mathematical Foundations of Interacting Multi-Particle Systems for Optimization

Konstantin Riedl

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Gero Friesecke

Prüfende der Dissertation: 1. Prof. Dr. Massimo Fornasier
2. Prof. Dr. Michael Herty
3. Prof. Dr. Nicolás García Trillos

Die Dissertation wurde am 10.04.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 02.09.2024 angenommen.

Abstract

Interacting multi-particle systems are of paramount importance in and beyond applied mathematics, with far-reaching impact across a variety of scientific disciplines.

This dissertation lays mathematical foundations for the numerical analysis of such systems in the setting of nonconvex nonsmooth optimization, itself a topic of fundamental interest throughout science and engineering. While systems of interacting particles are, due to their tremendous empirical success, broad spectrum of applicability, and ease of handling, widely used in practice, their rigorous theoretical analysis largely remained elusive. Given the necessity for capable, reliable, and robust algorithms that come with informative and solid convergence guarantees, a mathematical analysis framework for these methods is indispensable. We cover algorithms for classical global optimization problems in high dimensions as well as saddle point or so-called minimax problems. Our established analytical framework is flexible and versatile enough to be adapted to an even broader class of numerical methods. Furthermore, we discover a surprising, yet largely unexplored and unexploited link between the derivative-free and the gradient-based world in optimization.

The central observations and core contributions of this dissertation build upon theoretical insights obtained by taking a mean-field perspective, which, by alleviating original complexities of the problem, allows us to understand, unveil, and distill the internal mechanisms responsible for empirically observed successes. These findings, moreover, enable us to go beyond the investigated large particle regime and infer properties of the associated interacting multi-agent systems of practical interest.

Zusammenfassung

Interagierende Mehrteilchensysteme sind von herausragender Bedeutung in und jenseits der angewandten Mathematik, mit weitreichendem Einfluss auf eine Vielzahl von wissenschaftlichen Disziplinen.

Diese Dissertation legt mathematische Grundlagen für die numerische Analyse solcher Systeme im Rahmen der nichtkonvexen, nichtglatten Optimierung, einem Thema von fundamentaler Bedeutung in Wissenschaft und Technik. Obwohl Systeme von interagierenden Teilchen aufgrund ihrer enormen empirischen Erfolge, breiten Anwendbarkeit und einfachen Handhabung in der Praxis weit verbreitet sind, blieb ihre gründliche theoretische Analyse weitgehend aus. Angesichts der Notwendigkeit von leistungsfähigen, zuverlässigen und robusten Algorithmen mit aussagekräftigen und soliden Konvergenzgarantien ist ein mathematischer Rahmen für die Analyse jener Methoden unerlässlich. Wir behandeln Algorithmen für klassische globale hochdimensionale Optimierungsprobleme sowie Sattelpunkt- oder sogenannte Minimax-Probleme. Das von uns etablierte analytische Gerüst ist flexibel und vielseitig genug, um auf eine noch breitere Klasse numerischer Methoden angewandt zu werden. Darüber hinaus entdecken wir eine überraschende, bisher weitgehend unerforschte und ungenutzte Verbindung zwischen der ableitungsfreien und der gradientenbasierten Welt in der Optimierung.

Die zentralen Beobachtungen und Hauptbeiträge dieser Dissertation bauen auf theoretischen Erkenntnissen auf, die durch eine sogenannte Mittelfeldperspektive gewonnen werden, welche uns durch Abschwächung der ursprünglichen Komplexitäten des Problems ermöglicht, die internen Mechanismen, welche für empirisch beobachtete Erfolge verantwortlich sind, zu verstehen, enthüllen und herauszuarbeiten. Diese Erkenntnisse ermöglichen darüber hinaus, über das untersuchte Regime vieler Teilchen hinauszugehen und Eigenschaften der zugehörigen interagierenden Mehrteilchensysteme von praktischem Interesse abzuleiten.

Acknowledgements

When I look back over the last four years, the years of my PhD in Munich, I look back on a thrilling phase of my life. A phase wrapping up the decade of my life in Munich as a student. It has been a journey. It was supposed to be a journey.

I am grateful that I had this opportunity.

However, this journey would not have been the same if not for the amazing people I met along the way. The people who shaped me into who I am today, personally and academically.

Of them, first and foremost, I would like to express my deep gratitude to Massimo Fornasier for his generosity manifested in so many ways, for his guidance, for a unique amount of freedom and support blended with his caring and profound advice whenever needed and asked for. Thank you, Massimo, for your genuine support and sincere appreciation, which started way before I decided to start a PhD in your group and never ended since. It is a pleasure to work with you and learn from you.

I would also like to heartily thank the two other members of my examination committee, Michael Herty and Nicolás García Trillos, as well as Giuseppe Savaré, Lorenzo Pareschi, Elisabeth Ullmann, Felix Kraemer and Johannes Maly, for selflessly and generously supporting me over the last years and for your time to discuss about research, academia, and life in general.

Furthermore, I would like to thank two of my closest and earliest collaborators, Hui Huang and Timo Klock, for almost literally taking me by my hand when I took my first steps in academia. You did not just show me how to do research, but also how research should be done. Thank you, Hui. Thank you, Timo. But also, thank you, Massimo, for bringing us together.

Moreover, I would like to thank my PhD mentor, Daniel Kressner, for always listening, for offering his caring and wise advice as well as moral support, for looking after me.

I am incredibly fortunate to have made so many friends while being at work, be it in the office at our chair at TUM, be it when visiting and working from LMU or the city campus of TUM, be it at an event organized by IEAI, MDSI, or MCML, or be it on one of almost countless conferences, workshops, summer and winter schools, or research trips. Too many great people to be mentioned here, but let me mention the ones without whom the story of my PhD would have been a different one.

Johannes. For starting as my Master's thesis advisor years ago, accepting to be my unofficial PhD co-mentor, and eventually becoming a very dear friend.

Cristina, Alessandro, Hui, Tim, Giacomo, Ismael, Pascal. Vivian. Carolin. For being the greatest companions on this journey I could imagine and for becoming some of my closest friends.

Giacomo, Tim, Sixu, Nicolás, Yuhua, Leon, Jinniao, Lukang, Dohyeon. For making research fun and worth doing it, for being amazing colleagues and better friends despite

Acknowledgements

working all over the world. Moreover, thank you, Tim, for designing this thesis template and letting me use it, and thank you, Giacomo, for your generous and diligent proofreading.

Caterina, Ilaria, Alessio, Veronica. Maria Sofia. For making me feel welcome and home whenever I come to Italy.

Victoria, Philipp. Regina, Stefan, Lisa, Josef, Vitus, Marina, Carina, Daniel, Matthias, Ludwig. Katherina, Christoph. For taking the first steps with me in mathematics and in Munich, and for being a reason why I now call Munich my home.

Leonie. Kriti, Stefan. For caring for me, for being there whenever I need somebody to talk to, no matter what.

My Parents, Adrian, my Grandparents. For Everything.

Contents

Abstract	i
Acknowledgements	v
Preface	xi
Part I. Exposition	1
1. Introduction	3
2. Consensus-Based Optimization	7
2.1. The Dynamics of Consensus-Based Optimization	7
2.2. Variants and Applications of Consensus-Based Optimization	11
2.3. Code for Consensus-Based Optimization	16
3. Global Convergence of Consensus-Based Optimization and its Variants	19
3.1. A Global Convergence Analysis Framework for Consensus-Based Optimization	20
3.1.1. Global Convergence in Mean-Field Law	24
3.1.2. Mean-Field Approximation	36
3.1.3. Global Convergence in Probability: On Holistic Global Convergence Guarantees	49
3.2. Global Convergence of Consensus-Based Optimization with Truncated Noise	52
3.3. Global Convergence of Consensus-Based Optimization with Memory Effects and Gradient Information	55
4. Interpreting Consensus-Based Optimization as a Stochastic Relaxation of Gradient Descent	61
4.1. Consensus-Based Optimization Exhibits a Stochastic Gradient Descent-Like Behavior	61
4.2. From Consensus-Based Optimization to Consensus Hopping to Gradient Descent	66
5. Particle Swarm Optimization	69
5.1. The Dynamics of Particle Swarm Optimization	69

5.2.	Convergence of Particle Swarm Optimization to Global Minimizers . .	72
5.2.1.	Convergence of Particle Swarm Optimization with Memory Effects to Global Minimizers	72
5.2.2.	Convergence of Particle Swarm Optimization without Memory Effects to Global Minimizers	74
6.	Consensus-Based Optimization for Saddle Point Problems	77
6.1.	The Dynamics of Consensus-Based Optimization for Saddle Point Problems	77
6.2.	Convergence of Consensus-Based Optimization for Saddle Point Problems to Saddle Points	79
7.	Conclusions	83
	Bibliography	87
 Part II. Publications and Preprints		109
P1.	Consensus-Based Optimization Methods Converge Globally	111
P2.	Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law	161
P3.	Consensus-Based Optimization with Truncated Noise	195
P4.	Leveraging Memory Effects and Gradient Information in Consensus-Based Optimization: On Global Convergence in Mean-Field Law	237
P5.	CBX: Python and Julia Packages for Consensus-Based Interacting Particle Methods	277
P6.	Gradient is All You Need? How Consensus-Based Optimization can be Interpreted as a Stochastic Relaxation of Gradient Descent	293
P7.	On the Global Convergence of Particle Swarm Optimization Methods	339
P8.	Consensus-Based Optimization for Saddle Point Problems	391

Preface

This dissertation is structured into two intertwined parts. [Part I](#) provides an exposition of the topics and results discussed in more depth and presented in greater detail in the publications and preprints [[CBO-I](#); [CBO-II](#); [CBO-III](#); [CBO-IV](#); [CBX](#); [CBO&GD](#); [PSO](#); [CBO-SP](#)], which are collected and reprinted as [Papers P1](#) to [P8](#) in [Part II](#) of this document.

Part I: Exposition	Part II: Papers
Chapter 1: Introduction	—
Chapters 2 to 4: Consensus-Based Optimization (CBO): On Convergence, Variants and a Surprising Connection to Stochastic Gradient Descent	Papers P1 to P5 and Paper P6
Chapter 5: Particle Swarm Optimization (PSO)	Paper P7
Chapter 6: Approaching Saddle Point Problems with CBO	Paper P8
Chapter 7: Conclusions	—

Outline of the exposition. After highlighting the relevance of systems of interacting particles in both classical scenarios and the latest trends in optimization and artificial intelligence, we accentuate in [Chapter 1](#) the analytical power of mean-field perspectives when it comes to gaining a mathematical understanding of empirically observed phenomena.

In the core part of this dissertation, we lay *mathematical foundations* for the numerical analysis of *interacting multi-particle systems* for nonconvex nonsmooth *optimization* in high dimensions. While several of the central observations and main results of this work build upon theoretical insights obtained by taking a mean-field perspective, in many cases, these findings allow us to go beyond the investigated mean-field limits and infer properties of the associated multi-agent systems of practical interest. Actually, as we will see along the way, our journey will take us not just beyond the mean-field perspective, but also beyond interacting multi-particle methods and beyond optimization.

[Chapter 2](#) of this exposition is dedicated to the introduction of consensus-based optimization (CBO), a multi-particle metaheuristic derivative-free optimization method, which was originally proposed in [[Pin+17](#)]. A swarm of agents is employed to explore the domain and to form consensus about the location of the global minimizer by iteratively

computing a weighted average of all particles' positions, called consensus point, and consecutively evolving the swarm by each agent taking a step towards the consensus point while being subject to random noise featuring exploration. We discuss the motivations behind the design of the algorithm and provide an overview of its most relevant variants, applications and available code [CBX].

By summarizing the main contributions of [CBO-I; CBO-II] in the first section of Chapter 3, we present a versatile and flexible analysis framework for establishing global convergence guarantees for CBO methods. Our proof philosophy is as follows. Instead of directly investigating the microscopic particle system associated with the numerical algorithm, we first study the convergence behavior on the level of the continuous-time macroscopic agent density through the mean-field limit associated with the particle-based dynamics. By quantifying the convergence of the microscopic system to this mean-field limit as the number of employed particles grows in a second step, we eventually obtain a holistic convergence proof of the implementable CBO algorithm in form of a probabilistic global convergence result. The therewith developed framework has served as a template for proving the convergence of several variants and adaptations of CBO methods since then. By giving a brief overview of [CBO-III] and [CBO-IV] in the remaining sections of this chapter, two such variants of CBO are sketched together with their convergence analysis.

In the subsequent Chapter 4, leaving aside the mean-field analysis point of view for a moment, we shed light on the behavior of the CBO algorithm from a different angle by outlining the contributions of [CBO&GD]. By studying the trajectory of the consensus point of CBO, we observe that CBO exhibits a stochastic gradient descent-like behavior, which motivates the interpretation of CBO as a stochastic relaxation of gradient descent with a problem-tailored stochastic perturbation. The fundamental value of such link between CBO and stochastic gradient descent lies in the formerly established fact that CBO is provably globally convergent to global minimizers. Namely, on the one side, we offer a novel explanation for the success of stochastic relaxations of gradient descent and provide a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms, while, on the other side and contrary to the conventional wisdom for which zero-order methods ought to be inefficient or not to possess generalization abilities, we unveil an intrinsic gradient descent nature of such heuristics.

Returning to the mean-field analysis perspective from before but going beyond CBO, we turn in Chapter 5 towards the renowned particle swarm optimization method (PSO), which originated in the works [KE95; Ken97] and partially inspired the design of CBO. After concisely describing PSO, we summarize the work [PSO], where the convergence behavior of PSO to global minimizers is investigated under certain conditions of well-preparation of the hyperparameters and the initial datum by following the aforementioned philosophy and employing similar proof techniques.

Equipped with such a flexible analytical toolbox, we leave behind the quest of solving optimization problems and present in Chapter 6 consensus-based optimization for saddle point problems (CBO-SP) by providing an overview of the paper [CBO-SP], where CBO-SP is proposed, proven to converge to global Nash equilibria, and verified experimentally.

Chapter 7 wraps up the exposition of this dissertation.

List of publications and preprints. The following papers are part of this dissertation and reprinted as [Papers P1](#) to [P8](#) in [Part II](#) after the exposition.

- [CBO-I] M. Fornasier, T. Klock, and K. Riedl. “Consensus-Based Optimization Methods Converge Globally.” In: *arXiv preprint arXiv:2103.15130* (2021).
- [CBO-II] M. Fornasier, T. Klock, and K. Riedl. “Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law.” In: *Applications of Evolutionary Computation - 25th European Conference, EvoApplications 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings*. Ed. by J. L. J. Laredo, J. I. Hidalgo, and K. O. Babaagba. Vol. 13224. Lecture Notes in Computer Science. Springer, 2022, pp. 738–754.
- [CBO-III] M. Fornasier, P. Richtárik, K. Riedl, and L. Sun. “Consensus-Based Optimization with Truncated Noise.” In: *Eur. J. Appl. Math. (special issue “From integro-differential models to data-oriented approaches for emergent phenomena”)* (accepted 2024, to appear).
- [CBO-IV] K. Riedl. “Leveraging Memory Effects and Gradient Information in Consensus-Based Optimisation: On Global Convergence in Mean-Field Law.” In: *Eur. J. Appl. Math.* (accepted 2023, to appear), 32 pages.
- [CBX] R. Bailo, A. Barbaro, S. N. Gomes, K. Riedl, T. Roith, C. Totzeck, and U. Vaes. “CBX: Python and Julia packages for consensus-based interacting particle methods.” In: *arXiv preprint arXiv:2403.14470* (2024).
- [CBO&GD] K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier. “Gradient is All You Need?” In: *arXiv preprint arXiv:2306.09778* (2023).
- [PSO] H. Huang, J. Qiu, and K. Riedl. “On the Global Convergence of Particle Swarm Optimization Methods.” In: *Appl. Math. Optim.* 88.2 (2023), Paper No. 30, 44.
- [CBO-SP] H. Huang, J. Qiu, and K. Riedl. “Consensus-Based Optimization for Saddle Point Problems.” In: *SIAM J. Control Optim.* 62.2 (2024), pp. 1093–1121.

The accepted or published publications [CBO-II; CBO-IV; PSO; CBO-SP] are the core publications of this dissertation, whereas [CBO-III] is a further accepted publication and [CBO-I; CBX; CBO&GD] are further preprints contained in this dissertation.

Part I

Exposition

Chapter 1

Introduction

Systems of interacting agents or particles appear in a wide variety of scientific disciplines. They describe the physical movements in huge systems of atoms and molecules in molecular dynamics [KP90; HD18], the celestial motion [New87; BC61] of planets, stars, comets, and galaxies in astronomy, the collective behavior of large groups of people [CPT11; Alb+16], vehicles [CGP05; GP06; TK13], or animals [PE99; Cou+05; Sum10] in traffic flow or nature, the formation and dynamics of opinions [HK02; AO11] among humans in politics or among an obscure mixture of humans, institutions, and bots on social media, as well as the decentralized and distributed training [Dea+12; Ver+21; Kai+21] of nowadays large-scale machine learning models through a variety of digital devices ranging from pocket-size edge devices over servers to powerful compute clusters. In these scenarios, the particle interpretation arises naturally from real-world scenarios, and the systems are considered, modeled, and analyzed out of practical interest.

While the interaction rules can be seemingly simple, surprisingly plain and elegant in many cases, they fascinatingly enable the emergence of complex and often intelligent behavior, phenomena known as self-organization and swarm intelligence [Lor63; Man63; Bak96; VZ12]. For instance, a quartet of four fundamental forces, gravity, the weak force, electromagnetism, and the strong force, governs our universe and describes every physical interaction, from the creation of planets, solar systems, and galaxies, to the composition of matter at the level of quarks. On a cellular level, during biological ontogenesis, that is the development of an organism, embryonic cells exhibit coordinated behavior leading to the formation of spatio-temporal patterns [Isa12], thus the creation of life. In nature, all sorts of animals commit to flocking, herding, schooling, and milling behavior [Rey87], with one of the most startling examples being the herding of hundreds of thousands of gnus, zebras, and gazelles from the southern Serengeti in Tanzania to the lush green grasses of the Masai Mara in Kenya during the great migration. This list could be further expanded with several astonishing examples from all facets of nature, yet we conclude it here and refer the interested reader to the aforementioned books and references therein.

These intriguing capabilities, where individuals as a whole are more capable than they were on their own, have drawn researchers' attention toward specifically designing interacting multi-particle systems for a variety of purposes in different disciplines. In applied mathematics in particular, particle-based optimization algorithms [Hol75; KE95; Ken97; BFM97; Fog00; DB05; Yan14; Pin+17; Car+21; LTZ22; TZ24] look back on a long and successful history of being recognized as capable, reliable, and robust methods

that empirically achieve state-of-the-art performances on challenging global optimization problems, where the hardness is articulated through nonconvexity, nonsmoothness, and high dimensionality. Notable examples include evolution strategies [Sch95; SR95], evolutionary programming methods [BFM97; Fog00], genetic algorithms [Hol75], and particle swarm optimization (PSO) [KE95; Ken97]. Together with well-known optimization methods such as random search [Ras63], the Nelder-Mead simplex heuristic [NM65], the Metropolis-Hastings algorithm [Has70], and simulated annealing [AK89], such algorithms belong to the broad class of heuristics and metaheuristics [BR03; Bia+09; Tal09; Yan13; GP19]. Characteristically, these methods orchestrate an interplay between local improvement procedures and global strategies, combine deterministic and stochastic processes, leverage information exchange between multiple agents, to eventually design an efficient yet effective procedure for reliably and robustly searching the parameter space of an, in general, complicated objective function in search of a globally optimal solution.

With this philosophy and conceptual approach, the class of metaheuristics distinguishes itself substantially from and stands out against the classical paradigm prevalent in optimization [GMW20; Noc92; CST97; Fle01; BV04; NW06] by going beyond locality and focusing on a global exploration of the energy landscape, and consecutively, through communication, exploiting the gathered information. This contrasts algorithms such as gradient descent, the heavy ball method [Pol64; AGR00], Newton’s method and quasi-Newton methods [Bro67], or trust region methods [CGT00]. They either rely on local information about the objective function obtained through the evaluation of gradients or Hessians at the current iterate and consecutively invoke line search strategies to ensure a descent property, or directly restrain the search for a new iterate to a local neighborhood of the current iterate, the trusted region. Instead of being bound to converge locally as the formerly mentioned methods, metaheuristics aim at breaking these locality confinements by coupling stochasticity with the explorative power of a swarm of interacting particles and by featuring communication between the particles. Also stochastic variations of the aforementioned classical optimization algorithms, including stochastic gradient descent (SGD), AdaGrad [DHS11], RMSProp, and Adam [KB15], strive to become more capable of overcoming energy barriers of nonconvex functions by deploying randomness [Erm75; RT96; DM17; Chi22; ERY22], yet, with exploration through multiple agents in a swarm and communication between those being absent, they lack a key crucial and critical feature of many of the previously mentioned metaheuristics.

Despite the tremendous empirical success, broad spectrum of applicability, and widespread use of metaheuristics in practice throughout science and engineering, many of them, due to their inherent complexity and intricacies, lack proper mathematical foundations that could rigorously prove their robust convergence to global minimizers with explicit and quantitative rates under suitable assumptions. However, given the significance of solving complicated global optimization problems reliably and robustly throughout science and engineering, while having informative and solid convergence guarantees at disposal, this optimization paradigm is of substantial contemporary interest [Pin+17; Car+21; BBP22; LTZ22; TZ24], both from a practical and theoretical point of view.

The aforementioned difficulties, which arise when investigating stochastic systems with a large number of interacting particles, are attributed to three core and charac-

teristic features of metaheuristic methods. Firstly, a generally intricate dynamics and a nontrivial working principle of the algorithm. Secondly, the involved stochasticity. And lastly, but most crucially, a large number of particles paired with their interacting nature. To tackle such challenges, over the last decades, with origins in statistical mechanics [Bol77; Gib60], mean-field techniques have emerged as a powerful analytical tool and become a prominent and fruitful theoretical avenue when investigating large multi-particle systems. Examples include mean-field optimization [Orl85], the mean-field analysis of interacting particle systems in nature, sociology, and engineering [CCH14], mean-field games [LL07], mean-field optimal control [FS14; EHL19] as well as mean-field reinforcement learning [Yan+18]. Recently, an interpretation of neural networks as interacting particle systems with neurons being regarded as particles has allowed to gain a better understanding of the training process of wide shallow neural networks [RV18; RV22; MMN18; MMM19; CB18; SS20b; SS20a; FF22] as well as other architectures including deep networks [AOY19; SS22; Chi+22; NP23], ResNets [Din+22] and recurrent networks [LSS23]. In transformers [Vas+17], modeling tokens as particles [Lu+19; Dut+21] has paved the way for an analogy between the emergence of clusters in particle systems and next-token prediction [Ges+23b; Ges+23a] with the mean-field regime being natural in the setting of long-context understanding in the self-attention dynamics.

Contributions. In view of the thus far demonstrated paramount importance of interacting multi-particle systems in and beyond applied mathematics, paired with their far-reaching impact across a variety of scientific disciplines, we lay in this dissertation mathematical foundations for the numerical analysis of such systems in the setting of nonconvex nonsmooth optimization.

Based on the publications listed at the end of the foregoing preface, we present a mathematical analysis framework, that allows to derive rigorous quantitative estimates about the finite-time behavior for such numerical algorithms with explicit rates of convergence. This is of crucial interest and importance, and indispensable to warrant their applicability in particular in security-, privacy-, and fairness-sensitive applications. We cover algorithms for classical nonconvex global optimization problems in high dimensions as well as nonconvex-nonconcave saddle point or so-called minimax problems. Nevertheless, our established analytical techniques are flexible and versatile enough to be adapted and extended to an even broader class of methods.

Attributed to the difficulties of investigating large stochastic systems of interacting particles, the central observations and core contributions of this dissertation build upon theoretical insights obtained by taking a mean-field perspective, which, by alleviating original complexities of the problem, allows us to understand, unveil, and distill the internal mechanisms responsible for empirically observed successes. These findings, moreover, enable us to go beyond the investigated large particle regime and infer properties of the associated numerical algorithms of practical interest.

Furthermore, by viewing the interacting particle systems from a different angle, we discover a surprising, yet largely unexplored and unexploited link between the derivative-free and the gradient-based world in optimization.

Chapter 2

Consensus-Based Optimization

This chapter provides a gentle introduction to and overview of the field of consensus-based optimization (CBO), which is about a class of multi-agent metaheuristic derivative-free optimization methods with origins in the work [Pin+17]. In [Section 2.1](#), we describe the algorithm, explain its working principles, and elaborate on the motivations behind its design. The subsequent [Section 2.2](#) contains a survey of the most relevant variants and applications of CBO, demonstrating, in particular, the versatility and flexibility of the method. The chapter is wrapped up by [Section 2.3](#), where we survey available code for CBO.

2.1. The Dynamics of Consensus-Based Optimization¹

For the purpose of finding the global minimizer x^* of a potentially nonconvex, non-smooth, and high-dimensional objective function $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e., solving the unconstrained optimization problem

$$x^* \in \arg \min_{x \in \mathbb{R}^d} \mathcal{E}(x), \quad (2.1)$$

CBO methods employ a finite number of particles X^1, \dots, X^N to explore the domain and to form a consensus about the location of the minimizer x^* as time passes. In the spirit of metaheuristics [BFM97; BR03], the dynamics of the agents X^1, \dots, X^N of CBO are governed by two terms. A deterministic drift term drags each particle towards a weighted average of all agents' positions, referred to as the consensus point. Particles with a comparably low objective value, i.e., a presumably good position, are attributed a high weight in this average, whereas agents with a large objective value and, therefore, a worse position are assigned a lower weight and thus have less influence on the location of the consensus point. With this, the best position in the swarm is approximated, which serves as a proxy for the global minimizer x^* given the currently available information. The second term is stochastic in nature and randomly diffuses agents, thereby featuring the exploration of the energy landscape of the cost \mathcal{E} . The scaling of the noise is such that particles far from the consensus point, by being subject to larger noise, are able to explore larger regions of the domain, whereas agents in its proximity are affected by less

¹In this section, we follow [CBO-I, Section 1], [CBO-II, Section 1], and [CBO-IV, Section 1].

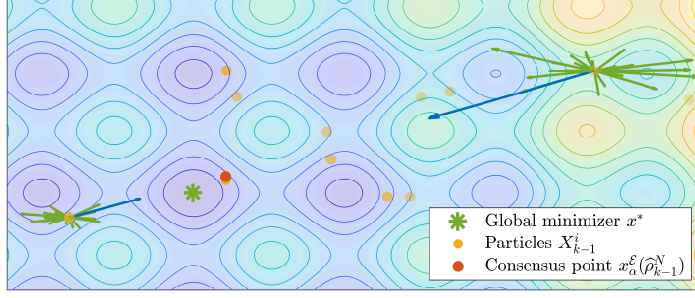


Figure 2.1: A visualization of the CBO dynamics (2.2) in the anisotropic diffusion setting (2.5). Particles X^1, \dots, X^N explore the energy landscape of the objective \mathcal{E} (the underlying function is the Rastrigin function) in search of the global minimizer x^* (green star). At time step k , the agents are located at positions $X_{k-1}^1, \dots, X_{k-1}^N$ (yellow dots). The consensus point $x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)$ (orange dot) is then computed as a weighted (visualized through color opacity of the particles) average of those positions. The dynamics of each agent (depicted exemplarily for just two agents) is governed by two terms. A consensus drift term (blue arrow) drags the respective particle towards the consensus point. A stochastic noise term (visualized by several green arrows depicting several possible realizations) injects randomness into the dynamics featuring the explorative nature of the algorithm.

noise and search the landscape of the objective function just locally. Before continuing with a more formal description of the method, a visualization of the dynamics is provided in Figure 2.1.

Given a finite time horizon $T > 0$ and a time discretization $0 < \Delta t < \dots < K\Delta t = T$ of $[0, T]$ with a suitable discrete time step size $\Delta t > 0$, we denote the position of the i th agent at time step k by $X_k^i \in \mathbb{R}^d$ and the empirical measure of all agents at time step k by $\hat{\rho}_k^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}$. For user-specified parameters $\alpha, \lambda, \sigma > 0$, the time-discrete evolution of the i th particle is given by the iterative update rule

$$X_k^i = X_{k-1}^i - \Delta t \lambda \left(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N) \right) + \sigma D \left(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N) \right) B_k^i, \quad (2.2)$$

where $((B_k^i)_{k=1, \dots, K})_{i=1, \dots, N}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}$. The system is complemented with independent initial data $(X_0^i)_{i=1, \dots, N}$, distributed according to a common initial law $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$. As mentioned in the informal description above, the updates in the evolution (2.2) consist of two terms, respectively. The first of which is the consensus drift, a deterministic drift towards the consensus point $x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)$, which is computed on the basis of the positions of the agents at time step $k-1$ and defined for a measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ according to

$$x_\alpha^\mathcal{E}(\varrho) := \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\varrho)}} d\varrho(x), \quad \text{with} \quad \omega_\alpha^\mathcal{E}(\bullet) := \exp(-\alpha \mathcal{E}(\bullet)). \quad (2.3)$$

That is, in the setting of the empirical measure $\widehat{\rho}_{k-1}^N$ and thus in the case of finitely many particles, we have $x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^N) = \sum_{i=1}^N X_{k-1}^i \omega_\alpha^\mathcal{E}(X_{k-1}^i) / \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_{k-1}^j)$, being the weighted average of the agents' positions at time step $k-1$. The choice of the weights in (2.3) is inspired by the Gibbs measure [Gib60], which has its origins in the Boltzmann distribution [Bol77] from statistical mechanics. The parameter α can, therefore, be interpreted as an inverse temperature. Mathematically, it is founded on the well-known Laplace principle [Hwa80; Mil06; DZ98], a classical result from large deviations theory, which states that, for any absolutely continuous probability distribution $\varrho \in \mathcal{P}(\mathbb{R}^d)$, we have

$$\lim_{\alpha \rightarrow \infty} \left(-\frac{1}{\alpha} \log \left(\int \omega_\alpha^\mathcal{E}(x) d\varrho(x) \right) \right) = \inf_{x \in \text{supp}(\varrho)} \mathcal{E}(x). \quad (2.4)$$

This justifies the interpretation of the consensus point $x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^N)$ as an approximation of $\arg \min_{i=1, \dots, N} \mathcal{E}(X_{k-1}^i)$, which improves as $\alpha \rightarrow \infty$, provided the minimizer uniquely exists. The second term in (2.2), stochastic in nature, encodes the diffusion mechanism of the method. It injects randomness into the dynamics, thereby featuring the explorative character of the algorithm. The two classically employed diffusion types are isotropic [Pin+17; Car+18; CBO-I] and anisotropic [Car+21; CBO-II] diffusion with

$$D(\bullet) = \begin{cases} \|\bullet\|_2 \text{Id}, & \text{for isotropic diffusion,} \\ \text{diag}(\bullet), & \text{for anisotropic diffusion,} \end{cases} \quad (2.5)$$

where $\text{diag} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ denotes the matrix-valued operator mapping a vector onto a diagonal matrix with the vector as its diagonal. Intuitively, scaling by $\|X_{k-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^N)\|_2$ or $\text{diag}(X_{k-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^N))$, respectively, encourages agents far from the consensus point to explore larger regions of the domain, whereas particles close to the consensus point try to enhance their position only locally. The scaling is furthermore essential to eventually deactivate the diffusive and explorative nature of the dynamics and to achieve consensus among the individual agents. Compared to isotropic diffusion, the coordinate-dependent scaling of anisotropic noise has proven to be more suitable for high-dimensional optimization problems [Car+21; CBO-II].

Motivation and inspiration.² The conceptual design of CBO is inspired and influenced by the renowned and well-known particle swarm optimization method (PSO) [KE95; Ken97]. In contrast to it, however, CBO was designed by the authors of [Pin+17] specifically and carefully to be amenable to a rigorous mathematical convergence analysis, which will be the focus of Chapter 3. It is worth mentioning, though, that the analytical techniques developed for and the insights gained from CBO can be transferred to suitable formulations of PSO, as we discuss in more detail in Chapter 5.

Like CBO, PSO methods follow a population-based paradigm to globally solve problems of the form (2.1). As per the description and derivations of [GP21], each particle in PSO is represented by a triplet $(X^{\text{PSO},i}, Y^{\text{PSO},i}, V^{\text{PSO},i}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, $i = 1, \dots, N$, consisting of a position, a personal (historical) best position, and a velocity, respectively.

²In this paragraph, we follow [GP21, Section 2].

With user-specified parameters $c_1, c_2 > 0$, the iterative update rule of the original PSO method reads

$$\begin{aligned} X_k^{\text{PSO},i} &= X_{k-1}^{\text{PSO},i} + \Delta t V_k^{\text{PSO},i}, \\ V_k^{\text{PSO},i} &= V_{k-1}^{\text{PSO},i} - \Delta t c_1 D \left(X_{k-1}^{\text{PSO},i} - Y_{k-1}^{\text{PSO},i} \right) R_{1,k}^i - \Delta t c_2 D \left(X_{k-1}^{\text{PSO},i} - y_{k-1}^{\text{PSO}} \right) R_{2,k}^i, \end{aligned} \quad (2.6)$$

where $((R_{1,k}^i)_{k=1,\dots,K})_{i=1,\dots,N}$ and $((R_{2,k}^i)_{k=1,\dots,K})_{i=1,\dots,N}$ are independent, identically distributed random vectors in \mathbb{R}^d with entries sampled uniformly from $[0, 1]$. $Y_{k-1}^{\text{PSO},i}$ denotes the personal (historical) best position of the i th agent, which it has seen up to time step $k-1$, explaining why it is also referred to as the memory of the associated particle $X^{\text{PSO},i}$. More formally, we set $Y_0^{\text{PSO},i} = X_0^{\text{PSO},i}$ and

$$Y_{k-1}^{\text{PSO},i} = \begin{cases} X_{k-1}^{\text{PSO},i}, & \text{if } \mathcal{E}(X_{k-1}^{\text{PSO},i}) < \mathcal{E}(Y_{k-2}^{\text{PSO},i}), \\ Y_{k-2}^{\text{PSO},i}, & \text{else,} \end{cases} \quad (2.7)$$

for $k \geq 2$. It is easy to observe that $Y_{k-1}^{\text{PSO},i} = \arg \min_{\ell=0,\dots,k-1} \mathcal{E}(X_\ell^{\text{PSO},i})$, where the position with the smallest index ℓ is chosen in case the objective attains the same value at multiple positions $X_\ell^{\text{PSO},i}$, $\ell = 0, \dots, k-1$. The quantity y_{k-1}^{PSO} , on the other hand, denotes the global best position of the swarm and is defined as

$$y_{k-1}^{\text{PSO}} = \arg \min \left\{ \mathcal{E}(Y_{k-1}^{\text{PSO},1}), \dots, \mathcal{E}(Y_{k-1}^{\text{PSO},N}), \mathcal{E}(y_{k-2}^{\text{PSO}}) \right\}. \quad (2.8)$$

In order to draw analogies to CBO, notice, that (2.6) can be reformulated equivalently as

$$\begin{aligned} X_k^{\text{PSO},i} &= X_{k-1}^{\text{PSO},i} + \Delta t V_k^{\text{PSO},i}, \\ V_k^{\text{PSO},i} &= V_{k-1}^{\text{PSO},i} - \Delta t \frac{c_1}{2} \left(X_{k-1}^{\text{PSO},i} - Y_{k-1}^{\text{PSO},i} \right) - \Delta t \frac{c_2}{2} \left(X_{k-1}^{\text{PSO},i} - y_{k-1}^{\text{PSO}} \right) \\ &\quad - \Delta t \frac{c_1}{2} D \left(X_{k-1}^{\text{PSO},i} - Y_{k-1}^{\text{PSO},i} \right) \tilde{R}_{1,k}^i - \Delta t \frac{c_2}{2} D \left(X_{k-1}^{\text{PSO},i} - y_{k-1}^{\text{PSO}} \right) \tilde{R}_{2,k}^i, \end{aligned} \quad (2.9)$$

where $((\tilde{R}_{1,k}^i)_{k=1,\dots,K})_{i=1,\dots,N}$ and $((\tilde{R}_{2,k}^i)_{k=1,\dots,K})_{i=1,\dots,N}$ are now independent, identically distributed random vectors with entries sampled uniformly from $[-1/2, 1/2]$, i.e., they in particular have zero mean.

Leaving aside the personal (historical) best positions of the particles for the moment, allowing to tune the deterministic and stochastic terms separately, changing the random vectors $\tilde{R}_{1,k}^i$ and $\tilde{R}_{2,k}^i$ to Gaussian random vectors, and regarding the consensus point (2.3) of CBO as an approximation of the global best position y_{k-1}^{PSO} , the PSO dynamics (2.9) resembles a second-order version of CBO. In fact, for the modified PSO optimizer proposed in [SE98], which introduces an inertia weight in the velocity update, this is made rigorous in [GP21; CHQ22] by proving that CBO can be derived from PSO in the zero-inertia limit.

Besides the rather apparent relation between CBO and PSO, the authors of [BP23] connect yet another class of metaheuristics, namely genetic algorithms, to CBO, see also [AFT23].

2.2. Variants and Applications of Consensus-Based Optimization³

Without claiming to exhaustively cover the full literature of the field of CBO, let us provide in what follows a brief overview of the most important variants and applications of CBO.

CBO with deactivatable consensus drift. The original version of CBO, put forth by the authors of [Pin+17], differs slightly from (2.2) by including an additional univariate function $H : \mathbb{R} \rightarrow [0, 1]$ in the drift term which, e.g., by setting $H(z) \approx \mathbb{1}_{z \geq 0}$, can be used to deactivate the consensus drift for those agents, whose position w.r.t. the objective function \mathcal{E} is better than the one of the consensus point, i.e., if $\mathcal{E}(X_{k-1}^i) < \mathcal{E}(x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N))$. More rigorously, their iterative update rule reads

$$\begin{aligned} X_k^i &= X_{k-1}^i - \Delta t \lambda \left(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N) \right) H \left(\mathcal{E}(X_{k-1}^i) - \mathcal{E}(x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)) \right) \\ &\quad + \sigma D \left(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N) \right) B_k^i. \end{aligned} \quad (2.10)$$

In the setting of isotropic diffusion, the simplified scheme with $H \equiv 1$ is analyzed⁴ mathematically in [Pin+17; Car+18], while in [CBO-I] we investigate both models, $H \equiv 1$ as well as $H \not\equiv 1$ satisfying $H(z) = 1$ whenever $z \geq 0$. The anisotropic diffusion model is studied in [Car+21] and [CBO-II] for $H \equiv 1$, see also Section 3.1. The theory of the latter can be extended to the case $H \not\equiv 1$ analogously to [CBO-I].

The works [HJK20; HJK21] directly investigate the time-discrete system (2.10), also for the case $H \equiv 1$, i.e., (2.2), but provided that the same random vector is used for all agents in the noise term, i.e., $(B_k^i)_{k=1, \dots, K} = (B_k)_{k=1, \dots, K}$ for all $i = 1, \dots, N$. Such a choice, however, leads to a less explorative dynamics.

CBO with drift to best particle instead of weighted mean. The authors of [BHW24] replace in the discrete scheme (2.2) the weighted mean $x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)$ as defined in (2.3) with its limit for $\alpha \rightarrow \infty$, i.e., $\arg \min_{i=1, \dots, N} \mathcal{E}(X_{k-1}^i)$. This goes back to the original implementation of the globally best position of the swarm in PSO, cf. (2.8).

CBO with truncated noise. In the work [CBO-III], which we explore in more detail in Section 3.2, we propose a variant of CBO that incorporates truncated noise in place of the classical noise term in order to enhance the well-behavedness of the statistics of the law of the dynamics. This yields improved convergence performance, allowing, in particular, for wider flexibility in choosing the noise parameter σ of the method, which, in turn, enables a more effective exploration of the energy landscape.

³In this section, we follow and extend [CBO-IV, Section 1, Paragraph 1].

⁴To be precise, [Pin+17; Car+18] as well as [Car+21; CBO-II] investigate the with (2.2) or (2.10) associated time-continuous dynamics from a mean-field perspective. Only, [CBO-I] provides a global convergence statement about the implementable schemes (2.2) and (2.10). This is the topic of Section 3.1. The results of [CBO-I], however, can be adapted straightforwardly to the anisotropic noise setting of [CBO-II] as done in Section 3.1.

CBO driven by jump-diffusion processes. With the aim of improving the explorative capabilities of the CBO dynamics as well, the paper [KST23] suggests and investigates the usage of discretized jump-diffusion processes in addition to the standard diffusion prevalent in CBO.

CBO with memory effects. Taking inspiration from the personal (historical) best features of particles, which are typical in the literature of PSO and which we described at the end of the preceding section, the authors of [GP21; TW20] proposed two different models for including such mechanisms in the CBO dynamics in order to increase the capabilities of the model. In [CBO-IV], which we elaborate on in Section 3.3, we study the convergence behavior of the CBO variant introduced in [GP21]. Also [BGP23] investigates, both theoretically and numerically, the advantages of memory mechanisms in this model, integrating further a random selection strategy to the dynamics, which leads to increased efficiency of the implementation.

CBO leveraging gradient information. For objective functions \mathcal{E} that have \mathcal{C}^1 regularity, the work [CBO-IV], which we investigate more closely in Section 3.3, extends the standard CBO dynamics (2.2) by appending a local gradient drift term for each particle. This allows to benefit from efficient local gradient improvements while retaining the capability of CBO to detect the basins of attraction of global minimizers of nonconvex objectives.

Ideologically similarly, the authors of [STW23] propose the exploitation of on-the-fly extracted higher-order differential information through inferred gradients based on point evaluations of the objective function during the CBO dynamics.

CBO with momentum. Taking inspiration from the advantages of stochastic gradient descent with momentum or Adam [KB15] over plain stochastic gradient descent, [KB15] presents a variant of CBO, called Adam-CBO, which brings adaptive momentum estimation to CBO.

Constrained CBO. In order to solve general constrained optimization problems with CBO, the works [CTV23; BHP23b] recast the constrained problem into a penalized (unconstrained) problem in order to be able to apply CBO in form of (2.2) to the modulated objective function which now includes a contribution from the constraint. While this does not explicitly constrain the dynamics to the feasible set, the particles are attracted towards the feasible set by means of the penalization term incorporated in the objective. In order to reinforce this behavior, [CTV23] introduces an additional relaxation drift towards the constraint manifold, which is determined in their case through equality and inequality constraints. The authors of [BHP23b], on the other hand, consider general feasible sets with merely a boundary of zero measure. Moreover, in order to employ exact penalization, they develop an iterative strategy that successively updates the penalization parameter depending on the violation of the constraints.

The work [Bae+22] follows a different strategy to take care of possible constraints by proposing a predictor-corrector-type CBO method, which projects the particles at the end of each time step onto the feasible set. For this purpose, the feasible set is required to be convex.

CBO on compact hypersurfaces. In contrast to the two first-mentioned papers in the preceding paragraph, the line of works [For+20; For+21; For+22] explicitly constrains the CBO dynamics to the feasible set by designing the method in a way such that it intrinsically remains in the respective set. This is done for hypersurfaces Γ , including, for instance, the sphere and the torus.⁵ For instance, for the sphere \mathbb{S}^{d-1} , the authors propose the iterative scheme

$$\begin{aligned}\tilde{X}_k^i &= X_{k-1}^i + \Delta t \lambda \mathbf{P}_{X_{k-1}^i} \left(x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N) \right) \\ &\quad + \sigma D \left(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N) \right) \mathbf{P}_{X_{k-1}^i} (B_k^i) + \text{noise correction term}, \quad (2.11) \\ X_k^i &= \mathbf{P}_{\mathbb{S}^{d-1}} \left(\tilde{X}_k^i \right),\end{aligned}$$

where the operators $\mathbf{P}_{\mathbb{S}^{d-1}}$ and \mathbf{P}_x , respectively, denote the projections onto the sphere \mathbb{S}^{d-1} and its tangent space $T_x \mathbb{S}^{d-1}$ at $x \in \mathbb{S}^{d-1}$, i.e., $\mathbf{P}_x = \text{Id} - xx^T$. Notice, that the deterministic drift term in the first line of (2.11) mimics the consensus drift term since $\mathbf{P}_{X_{k-1}^i}(X_{k-1}^i) = 0$ and thus $\mathbf{P}_{X_{k-1}^i}(x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)) = -\mathbf{P}_{X_{k-1}^i}(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N))$

[Kim+20] considers the special case of constrained optimization problems over the Stiefel manifold.

CBO on graphs. The thesis [Vli20] investigates an adaptation of CBO to a different topology, namely graphs $\mathcal{G} = (V, E)$. This requires to ensure that the employed particles move over the fixed set of vertices V by using only the edges E of the graph. To this end, while the consensus point is computed in the familiar fashion, the stochasticity of the standard CBO method is converted into a probability distribution over the neighborhood of each particle and determines the likelihood of which edge is taken by the particle in the next step.

Polarized CBO. Turning back to unconstrained optimization problems, one often encounters objective functions with several global minimizers. With the standard CBO dynamics being designed to reach consensus in the limit at a unique point, a modification is necessary to make CBO capable of detecting multiple global minimizers in parallel. To this end, the authors of [BWR22] propose to polarize the standard CBO dynamics by replacing the common consensus point with one weighted mean per particle, which

⁵While devising a CBO method which is intrinsic to certain hypersurfaces is of interest, one central contribution of [For+20; For+21; For+22] was to provide the first-ever holistic convergence proof of any CBO method to global minimizers by proving a mean-field approximation result and thus being able to obtain a convergence statement beyond the mean-field limit. Although the authors heavily use the fact of being constrained to a compact set, some of the techniques developed there, made possible the work [CBO-I], which we discuss in Section 3.1.

attributes more weight to nearby particles. This is realized by introducing a localization kernel $\mathbf{k} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_0^+$ in the computation of the weighted mean, taking inspiration from bounded confidence models of opinion dynamics [Def+00; HK02; For+05; GGL12]. A suitably localized weighted mean is given by

$$x_{\alpha, \mathbf{k}}^{\mathcal{E}}(\varrho, x) := \int z \frac{\mathbf{k}(x, z) \omega_{\alpha}^{\mathcal{E}}(z)}{\|\mathbf{k}(\bullet, z) \omega_{\alpha}^{\mathcal{E}}(\bullet)\|_{L^1(\varrho)}} d\varrho(z), \quad \text{with } \omega_{\alpha}^{\mathcal{E}}(\bullet) := \exp(-\alpha \mathcal{E}(\bullet)) \quad (2.12)$$

and, for instance, with the finite range interaction kernel $\mathbf{k}(x, z) = \mathbb{1}_{\|x-z\|_2 \leq \kappa}(x, z)$ or a Gaussian kernel $\mathbf{k}(x, z) = \exp(-\|x-z\|_2^2 / (2\kappa^2))$, where κ is referred to as confidence level. The time-discrete evolution of the i th particle is now given by the iterative update rule

$$X_k^i = X_{k-1}^i - \Delta t \lambda \left(X_{k-1}^i - x_{\alpha, \mathbf{k}}^{\mathcal{E}}(\widehat{\rho}_{k-1}^N, X_{k-1}^i) \right) + \sigma D \left(X_{k-1}^i - x_{\alpha, \mathbf{k}}^{\mathcal{E}}(\widehat{\rho}_{k-1}^N, X_{k-1}^i) \right) B_k^i, \quad (2.13)$$

which makes the dynamics of the i th particle depend mainly on spatially close particles due to the computation of the weighted mean in (2.12). This polarizes the optimization dynamics, featuring the clustering capability of the polarized CBO method and permitting the finding of multiple global minimizers at the same time.

In [Brü21], where this concept is referred to as CBO with finite range interaction, a similar direction is pursued to enable clustering in the CBO dynamics and thereby making the method capable of finding multiple global minimizers.

However, both works lack a rigorous mathematical analysis in a general nonconvex setting. Such analysis is conducted by the authors of [FS24], where the associated underlying degenerate nonlinear nonlocal Fokker-Planck equation is analyzed by adopting the techniques of [CBO-I] to the more involved setting of polarized CBO. This requires the use of a kernel \mathbf{k} , which allows for an adaptive scaling.

CBO for multi-objective optimization problems. The line of works [BHP22; BHP23a; Bor23] brings CBO to the problem class of multi-objective optimizations [Jah04; HM79; PŽŽ17; Eic21], where several functions are optimized simultaneously with the aim of finding an optimal solution set, known as the Pareto front. The author's approach is founded on a scalarization strategy which reduces the multi-objective problem to an infinite number of single-objective scalar sub-problems. By devising the interacting particle in a way that each particle optimizes a different of these parametrized sub-problems while still communicating to exchange information about the objective functions, the entire Pareto front can be approximated with a single run of the algorithm. To ensure that the particles are well-distributed over the computed front, CBO is coupled with a binary repulsive dynamics in the parameter space.

A different strategy is pursued in [KST24], where a multi-swarm CBO algorithm is employed with each swarm solving one of the scalarized problems. Again, to allow for a diverse approximation of the front, the swarms interact through adaptive scalarization weights.

CBO for distributed training and federated learning. In many applications of modern machine learning and computer science in general, distributed training or distributed optimization [RN04; Yan+19] is of considerable practical interest. To make the training process scalable or because of the natural architecture of the task, both workload and data are shared across several devices, making communication between them crucial and critical. In federated learning [McM+17; Kai+21] scenarios, the training procedure is moreover subject to data privacy considerations and characterized by data heterogeneity.

To formalize this problem type, let us consider N agents or devices with positions X^i representing the parameters of the neural network each agent wants to train. While each agent has access only to its private data, captured by its own loss function \mathcal{E}_i , it would like to benefit from the information and the data of others, if they are relevant to the particular agent. For instance, in the specific instance of clustered federated learning [Gho+20; SMS21], an unknown group structure is assumed to underly the individuals possessing the computational resources and training data. More formally, in the case of \aleph groups, we assume that the objective function \mathcal{E}_i of each agent is of the form $\mathcal{E}_i = \tilde{\mathcal{E}}_j + n_i$ for some $j \in \{1, \dots, \aleph\}$. Here, $\tilde{\mathcal{E}}_j$ can be regarded as the loss of the j th group and n_i models the deviation of the i th agent’s loss from the one of the group. With n_i being typically small, individuals belonging to the same group are thus assumed to have similar objectives. Despite the individual agents being oblivious to their group membership, each group eventually intends to collaboratively train a different model.

In the more idealized setting, where $n_i = 0$ for all agents, the authors of [Car+23] propose a consensus-type optimizer, dubbed FedCBO, which is provably capable of training machine learning models in the data privacy-sensitive setting of federated learning with heterogeneity in the data as captured by the existence of multiple groups. FedCBO achieves this by devising an interacting particle system, oblivious to group membership, where each individual, after receiving the models from the other devices, computes the consensus point w.r.t. its private loss function and then runs the dynamics

$$X_k^i = X_{k-1}^i - \Delta t \lambda \left(X_{k-1}^i - x_\alpha^{\mathcal{E}_i}(\hat{\rho}_{k-1}^N) \right) + \sigma D \left(X_{k-1}^i - x_\alpha^{\mathcal{E}_i}(\hat{\rho}_{k-1}^N) \right) B_k^i. \quad (2.14)$$

With this, agents or devices from the same group, thus having the same or, in the non-idealized setting, similar loss functions benefit from the models of each other when computing the consensus point, while models from different groups are typically filtered by being attributed less weight due to a significant difference in the loss functions. To speed up the training and limit communication, individuals train their models locally with gradient steps between two communication rounds, i.e., the iterative update rule (2.14) is appended by a local gradient as suggested in [CBO-IV].

Consensus-based sampling. Attributed to a remarkable connection between sampling and the field of optimization [Che23], succinctly captured by the motto “sampling is optimization in the space of measures” [Wib18], and mathematically founded on the seminal work [JKO98], a sampling analog of CBO is natural to be introduced. By modifying the diffusion term of CBO to include a weighted sample covariance preventing the

collapse of the particle ensemble to full consensus, [Car+22] proposes consensus-based sampling (CBS), which is designed to converge to a Gaussian approximation of the distribution $\frac{1}{Z} \exp(-\mathcal{E})$ with Z denoting the normalization factor [GS90; RC04; Liu01]. CBS can be utilized to generate approximate samples from a given target distribution by running the interacting particle system. It can, moreover, be employed in optimization mode to solve problems of the form (2.1).

In analogy to the aforementioned polarized CBO variant, the authors of [BWR22] furthermore devise a CBS variant, which allows sampling from distributions with multiple modes.

CBO for saddle-point problems and multiplayer games. As we briefly address in Chapter 6, a CBO-type algorithm, called CBO-SP, has been developed for saddle point problems [Nas50; BGL05] in [CBO-SP]. It employs a group of interacting particles, which perform a minimization over one variable and a maximization over the other.

The work [CHQ23] extends this paradigm to multiplayer games. Both methods reliably identify global Nash equilibria for nonconvex problems.

Applications of CBO. In the collection of formerly referenced works, CBO has demonstrated to be a valuable and versatile method for a wide scope of applications reaching from phase retrieval, robust subspace detection, compressed sensing, or image segmentation problems in signal processing [For+21; For+22; CBO-IV; BGP23] to the training of neural networks for image classification in machine learning [Car+21; CBO-II; CBO-IV; BGP23; PSO], even in the data privacy-sensitive setting of federated learning [Car+23]. It has been furthermore employed to solve a wide range of linear and non-linear ordinary differential equations [Nik22], to approximate low-frequency functions in the presence of high-frequency noise and to the task of solving PDEs with low-regularity solutions [CJL22], in asset allocation in finance [Bae+22], for oligopoly games with several goods in economics [CHQ23], optimal actuator and control design [Kal+24], rare event estimation in uncertainty quantification [APU23], or to simulate chemical reactions [BH23] by coupling CBO with model predictive control strategies.

2.3. Code for Consensus-Based Optimization⁶

MATLAB implementations of CBO and some of its variants [CBO-I; CBO-II; CBO-III; CBO-IV] can be found in the GitHub repository [CBOGlobalConvergenceAnalysis](#) as well as the repositories of the cited references, including in particular [AM-CBO](#) for CBO for multi-objective optimization [BHP22; BHP23a; Bor23] and [KV-CBO](#) for CBO on compact hypersurfaces [For+20; For+21; For+22]. Code for CBO for saddle-point problems [CBO-SP] can be found in [CBOSaddlePoints](#). The repository [PSOAnalysis](#) provides code for PSO [PSO].

With the packages [CBXpy](#) and [CBX.jl](#) we provide unified Python and Julia implementations of several consensus-based interacting particle methods [CBX]. They offer a

⁶In this section, we follow [CBX].

lightweight, easy-to-understand, -use and -extend implementation of CBO together with several of its variants, including CBO with mini-batch ideas [Car+21; CBO-II], CBO with restart [Car+21; CBO-II], a cooling strategy of the parameters [For+21; CBO-II], polarized CBO [BWR22], CBO with memory effects [GP21; CBO-IV], PSO [GP21; Gra+23; PSO], CBS [Car+22], and more to come. The defined structures and hierarchies in the code ensure a usage experience similar to optimizer classes in `scikit-opt` and `PyTorch` [Pas+19]. Numerous utilities, like performance evaluation or plotting routines tailored to CBO methods are provided. The code of these packages builds upon the repositories `polarcbo`, where polarized CBO [BWR22] is implemented, as well as `cbo-in-python`, and `Consensus.jl`, respectively. `FedCBO` moreover implements Fed-CBO from [Car+23]. `CBO-multiplayer` provides Python code for CBO for multiplayer games [CHQ23].

Chapter 3

Global Convergence of Consensus-Based Optimization and its Variants

In this chapter, we turn towards the first core contribution of this dissertation, the global convergence analysis of CBO methods. After outlining in [Section 3.1](#) the philosophy behind our analysis strategy, which takes at its heart a mean-field perspective, we revisit and provide in the first part of [Section 3.1.1](#) an overview of the works [[Pin+17](#); [Car+18](#); [Car+21](#)], where the idea of investigating the mean-field limit of CBO to understand the algorithm’s convergence and optimization behavior was first suggested and pursued. By establishing consensus formation of the mean-field limit of CBO, and consecutively showing that the found consensus lies close to a global minimizer provided certain well-preparedness conditions, the authors obtain the macroscopic convergence of CBO to global minimizers of the objective function under the aforementioned locality assumptions. Thereafter, we motivate and present the analytical framework put forth in our papers [[CBO-I](#); [CBO-II](#)]. While it is in a similar spirit as the preceding works, in particular, with a convergence analysis of the mean-field limit of CBO being a key aspect, it differs considerably in several aspects, which we elaborate on in more detail in [Sections 3.1.1](#) to [3.1.3](#). First of all, in the second part of [Section 3.1.1](#) we present a novel technique for proving global convergence to the minimizer in mean-field law, which is valid for a rich class of objective functions and is based on the analysis of the Wasserstein-2 distance between the law of the mean-field CBO dynamics and a Dirac measure located at the global minimizer of the objective function as well as a quantitative nonasymptotic Laplace principle. This unveils, in particular, that CBO performs a convexification of a large class of optimization problems as the number of optimizing agents goes to infinity. In order to leverage the result about convergence in mean-field law, [Section 3.1.2](#) establishes a probabilistic mean-field approximation that quantifies how well the microscopic interacting particle system approximates the macroscopic mean-field limit as the number of employed particles grows. Combining the statements of the two former sections in [Section 3.1.3](#) allows us to obtain probabilistic global convergence guarantees of the numerical CBO method ([2.2](#)). The chapter proceeds with [Sections 3.2](#) and [3.3](#), where we cover the works [[CBO-III](#)] and [[CBO-IV](#)], which provide global convergence results for two variants of CBO, namely CBO with truncated noise and CBO with memory effects and gradient information, by following the formerly established framework. With this,

we highlight the versatility and flexibility of our framework, which has been adopted by several other research groups for some of the CBO variants mentioned in [Section 2.2](#).

3.1. A Global Convergence Analysis Framework for Consensus-Based Optimization⁷

Analyzing directly the interacting particle system (2.2), i.e., investigating CBO on its microscopic scale, poses a challenging endeavor, for one thing due to the nonlinearity of the dynamics, its stochasticity, and a possibly large dimensionality d of the problem, but in particular as a consequence of the typically large number of particles N paired with their interacting nature. This introduces correlations between the individual stochastic processes that represent the involved agents, and requires the whole system to be considered and analyzed in its entirety, which may be intractable, both computationally and mathematically, due to the complexity. Analysis attempts in this direction have been undertaken in [\[HJK20; HJK21\]](#), however, under a substantially restrictive assumption on the model, namely that all particles use at each time step the identical random vector B_k^i , i.e., $(B_k^i)_{k=1,\dots,K} = (B_k)_{k=1,\dots,K}$ for all $i = 1, \dots, N$. Such choice, however, leads to a significantly less explorative and less capable dynamics.

The analytical difficulties mentioned at the beginning of this section hold identically for the continuous-time version of (2.2), which we investigate in place of its discrete-time counterpart in most of the parts of this dissertation out of analytical convenience. It is given by the system of stochastic differential equations (SDEs), expressed in Itô's form as⁸

$$dX_t^i = -\lambda \left(X_t^i - x_\alpha^\mathcal{E}(\widehat{\rho}_t^N) \right) dt + \sigma D \left(X_t^i - x_\alpha^\mathcal{E}(\widehat{\rho}_t^N) \right) dB_t^i, \quad (3.1)$$

where $((B_t^i)_{t \geq 0})_{i=1,\dots,N}$ are independent standard Brownian motions in \mathbb{R}^d . As custom from before, the system is complemented with the independent initial data $(X_0^i)_{i=1,\dots,N}$, distributed according to a common initial law $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$. Moreover, the consensus point $x_\alpha^\mathcal{E}$, which is now computed instantaneously from the particles' positions, is defined as in (2.3) and the measure $\widehat{\rho}_t^N$ denotes the empirical measure of the system (3.1), i.e., $\widehat{\rho}_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$. Equation (2.2) originates from a simple Euler-Maruyama time discretization⁹ [\[KP92; Pla99; Hig01\]](#) of (3.1), with standard approximation results being available in the literature. However, let us remark, that the ideas and techniques presented in the following sections and giving an overview of the works [\[CBO-I; CBO-II\]](#) can be transferred also to the discrete-time setting. In fact, the thesis [\[Bor24\]](#) pursues this direction in great detail.

⁷In this section, we follow and extend [\[CBO-I, Section 1\]](#) and [\[CBO-II, Sections 1 and 2\]](#) integrating parts of [\[CBO-IV, Section 2\]](#).

⁸Notice here, that we slightly abuse notations by using the same notation X_t^i for both (2.2) and (3.1). However, it will be clear from the context, to what we refer.

⁹The Euler-Maruyama method, being an extension of the explicit Euler method from ordinary to stochastic differential equations, is just one possible time discretization scheme. For alternatives see [\[KP92; Pla99; Hig01\]](#).

With the microscopic systems (2.2) and (3.1) being intricate to analyze, the authors of [Pin+17] have proposed to resort to investigating a macroscopic description of the dynamics. Instead of trying to capture and describe the trajectories of all particles individually, a statistical description in terms of a single particle distribution is sought. This concept is reminiscent of classical approaches in statistical mechanics [Bol77], where one is interested in the physical properties of the system as a whole rather than the behavior of its components. In the original example of thermodynamics, for instance, the number of electrons, atoms, molecules, or other constituents is enormous, making it impossible to utilize the knowledge about their atomic interaction principles to simulate or analyze the respective particle system as a whole and to then infer properties about typical macroscopic quantities of interest, like pressure, volume, and temperature. However, in most cases, the microscopically present complexities are not necessary to be captured in order to describe macroscopic phenomena.

The intuitive rationale is as follows. In a system with a vast number of particles or agents, one expects that, for any particle, the individual influence of any other particle disperses, resulting in an averaged influence of the ensemble rather than an interacting nature of the system. Heuristically, as the number of particles tends to infinity, they are expected to become statistically independent and behave the same, encouraging the description of the macroscopic dynamics in the large-particle limit in a statistical or probabilistic way. For CBO, the dynamics of such typical particle can be captured by the self-consistent SDE¹⁰

$$d\bar{X}_t = -\lambda \left(\bar{X}_t - x_\alpha^\mathcal{E}(\rho_t) \right) dt + \sigma D \left(\bar{X}_t - x_\alpha^\mathcal{E}(\rho_t) \right) dB_t, \quad (3.2)$$

where the statistical influence of the ensemble is embodied through $\rho_t = \text{Law}(\bar{X}_t)$, i.e., the statistical behavior of the typical particle itself. The SDE is complemented by the initial condition $\bar{X}_0 \sim \rho_0 \in \mathcal{P}(\mathbb{R}^d)$. An application of Itô's formula shows that the measure $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ with $\rho(t) = \rho_t = \text{Law}(\bar{X}_t)$ satisfies the nonlinear nonlocal¹¹ Fokker-Planck equation

$$\partial_t \rho_t = \lambda \text{div} \left(\left(x - x_\alpha^\mathcal{E}(\rho_t) \right) \rho_t \right) + \frac{\sigma^2}{2} \sum_{k=1}^d \partial_{kk} \left(D \left(x - x_\alpha^\mathcal{E}(\rho_t) \right)_{kk}^2 \rho_t \right) \quad (3.3)$$

in a weak sense (see Definition 3.1). This yields the desired macroscopic description of the CBO dynamics through a particle distribution ρ in terms of a partial differential equation (PDE).

Definition 3.1 (Weak solution, [CBO-II, Definition 1]). Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$, $T > 0$. We say $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ satisfies the Fokker-Planck equation (3.3) with initial condition ρ_0 in the weak sense in the time interval $[0, T]$, if for all $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and all

¹⁰This self-consistent SDE is sometimes also referred to as mono-particle process.

¹¹Both nonlinearity and nonlocality of the PDE are consequences of the definition of the consensus point (2.3).

$t \in (0, T)$ it holds

$$\begin{aligned} \frac{d}{dt} \int \phi(x) d\rho_t(x) &= -\lambda \int \sum_{k=1}^d (x - x_\alpha^\mathcal{E}(\rho_t))_k \partial_k \phi(x) d\rho_t(x) \\ &\quad + \frac{\sigma^2}{2} \int \sum_{k=1}^d D(x - x_\alpha^\mathcal{E}(\rho_t))_{kk}^2 \partial_{kk}^2 \phi(x) d\rho_t(x) \end{aligned} \quad (3.4)$$

and $\lim_{t \rightarrow 0} \rho_t = \rho_0$ pointwise.

Founded on the aforementioned philosophy, we coined in [CBO-I; CBO-II] the notion of convergence in mean-field law, given as in Definition 3.2.

Definition 3.2 (Convergence in mean-field law, [CBO-I, Definition 1]). Let $F, G : \mathcal{P}(\mathbb{R}^d) \otimes \mathbb{R}^d \rightarrow \mathbb{R}^d$ be two functions and consider for $i = 1, \dots, N$ the interacting system of SDEs expressed in Itô's form as

$$dX_t^i = F(\widehat{\rho}_t^N, X_t^i) dt + G(\widehat{\rho}_t^N, X_t^i) dB_t^i, \quad \text{where } \widehat{\rho}_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}, \text{ and } X_0^i \sim \rho_0. \quad (3.5)$$

We say that this SDE system converges in mean-field law to a point $\tilde{x} \in \mathbb{R}^d$ if all solutions of the self-consistent SDE

$$d\bar{X}_t = F(\rho_t, \bar{X}_t) dt + G(\rho_t, \bar{X}_t) dB_t, \quad \text{where } \rho_t = \text{Law}(\bar{X}_t), \text{ and } \bar{X}_0 \sim \rho_0, \quad (3.6)$$

satisfy $\lim_{t \rightarrow \infty} W_p(\rho_t, \delta_{\tilde{x}}) = 0$ for some Wasserstein- p distance W_p , $p \geq 1$.

Colloquially speaking, an interacting multi-particle system is said to converge in mean-field law, if the associated mean-field dynamics converges.

A closer look reveals, that the self-consistent dynamics (3.6) is derived from its interacting counterpart (3.5) by merely replacing the empirical measure $\widehat{\rho}_t^N$ with its own law ρ_t . In the setting of CBO, (3.6) corresponds to (3.2) and (3.5) to (3.1). While this derivation of the mean-field CBO dynamics (3.2) and thus (3.3) from (3.1) is in the spirit of the formerly elaborated on philosophy, it is purely formal. However, the so-called mean-field approximation, i.e., the question of how well the mean-field dynamics ρ is approximated by $\widehat{\rho}^N$ w.r.t. the number of particles N has been made rigorous through several works as outlined in detail in Section 3.1.2.

Such results substantiate and legitimate the analysis of the mean-field CBO dynamics (3.2) and (3.3) in lieu of the interacting particle system (3.1), thereby justifying the notion of convergence in mean-field law as defined in Definition 3.2. This is the focus of Section 3.1.1. Yet, after having gained insights into the behavior of the CBO dynamics on a macroscopic level, the mean-field approximation results presented in Section 3.1.2 allow to transfer the results to the microscopic regime as done in Section 3.1.3. Together

with classical results of numerical approximation of SDEs [Pla99], eventually, convergence guarantees for the implementable CBO scheme (2.2) are obtained in Section 3.1.3

The following tableau provides an overview of the structure of the remaining section and points to the central statements and technical tools.

Global convergence of CBO methods to the global minimizer (Section 3.1.3, in particular Theorem 3.19)		
Global convergence in mean-field law (Section 3.1.1, i.p. Theorem 3.6)	Probabilistic mean-field approximation (Section 3.1.2, i.p. Proposition 3.16)	Numerical approximation of SDEs ([Pla99; KP92])
Time-evolution inequalities for \mathcal{V} (Lemmas 3.10 and 3.11)	Probabilistic moment bounds (Lemma 3.15)	
Quantitative Laplace principle (Proposition 3.12)	Probabilistic stability estimate for $x_\alpha^\mathcal{E}$ (Lemma 3.17)	
A lower bound for mass around x^* (Proposition 3.13)	Probabilistic sampling estimate for $x_\alpha^\mathcal{E}$ (Lemma 3.18)	

Before continuing with the details about the mathematical convergence analysis, let us mention that under Assumption 3.3 the interacting particle system (3.1) as well as the mean-field dynamics (3.2) and (3.3) are well-posed in the sense of Hadamard [Had02], i.e., their respective solutions exist and are unique, see Theorem 3.4.

Assumption 3.3. Throughout we are interested in objective functions $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ with $\underline{\mathcal{E}} > -\infty$, for which

W1 there exist constants $C_1, C_2 > 0$ such that

$$|\mathcal{E}(x) - \mathcal{E}(x')| \leq C_1(1 + \|x\|_2 + \|x'\|_2)\|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d, \quad (3.7)$$

$$\mathcal{E}(x) - \underline{\mathcal{E}} \leq C_2(1 + \|x\|_2^2), \quad \text{for all } x \in \mathbb{R}^d, \quad (3.8)$$

W2 either $\bar{\mathcal{E}} := \sup_{x \in \mathbb{R}^d} \mathcal{E}(x) < \infty$, or there exist constants $C_3, C_4 > 0$ such that

$$\mathcal{E}(x) - \underline{\mathcal{E}} \geq C_3 \|x\|_2^2, \quad \text{for all } \|x\|_2 \geq C_4. \quad (3.9)$$

W1 requires that \mathcal{E} is locally Lipschitz-continuous with the Lipschitz constant being allowed to have linear growth. This entails in particular that the objective has at most quadratic growth at infinity as formulated explicitly in (3.8). Let us further remark that some papers assume $|\mathcal{E}(x) - \mathcal{E}(x')| \leq C_1(\|x\|_2 + \|x'\|_2)\|x - x'\|_2$ instead of (3.7), which is an unnecessary restriction since all theoretical considerations hold without technical difficulties immediately also for (3.7). A slightly more general set of assumptions is considered in [GHV23, Sections 2.2 and 4]. W2, on the other hand, assumes that \mathcal{E} also has at least quadratic growth in the farfield, i.e., overall it grows quadratically far away from x^* . Alternatively, \mathcal{E} may be bounded from above. Since the objective function \mathcal{E} can be usually modified for the purpose of analysis outside a sufficiently large region, these growth conditions are not really restrictive.

Under these assumptions, we have the following statement.

Theorem 3.4 (Well-posedness of CBO). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy W1–W2. Let $T > 0$ and $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then the following well-posedness statements hold.

- (i) For each $N \in \mathbb{N}$, there exists a unique strong solution $((X_t^i)_{t \in [0, T]})_{i=1, \dots, N}$ of the system of SDEs (3.1).
- (ii) There exists a unique nonlinear process $\bar{X} \in \mathcal{C}([0, T], \mathbb{R}^d)$ satisfying (3.2) in the strong sense. The with \bar{X} associated law $\rho = \text{Law}(\bar{X})$ has regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ and is a weak solution to the Fokker-Planck equation (3.3).

The proof follows verbatim the ones for the case of isotropic diffusion in (2.5), for which we refer the reader to [Car+18, Section 2] and [Car+18, Section 3].

We now turn to the global convergence analysis of CBO methods. Motivated by the former argumentation, let us start with investigating the mean-field perspective (3.2) and (3.3), respectively. The results are presented in the setting of anisotropic noise in (2.5), i.e., we first report on the paper [Car+21], before extending the work [CBO-II], where only the mean-field analysis for anisotropic CBO has been conducted. For CBO with isotropic diffusion analogous statements can be found in [Car+18] and [CBO-I], respectively.

3.1.1. Global Convergence in Mean-Field Law

For our first aim of investigating the global convergence behavior of CBO to a minimizer x^* of the objective function \mathcal{E} on the mean-field level, i.e., establishing for a weak solution $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ to (3.3) the convergence

$$\rho_t \rightharpoonup \delta_{x^*} \quad \text{as } t \rightarrow \infty \quad (3.10)$$

in some sense, two analytical frameworks are available in the literature. On the one side, the authors of [Car+18; Car+21] suggest a two-part approach, which first establishes consensus formation of the measure ρ_t in time by showing that the variance $\text{Var}(\rho_t)$ converges to 0 as $t \rightarrow \infty$, provided the initial distribution ρ_0 satisfies certain well-preparedness assumptions. Consecutively, by suitable choices of the parameters of

the dynamics, in particular a large value for α , it is ensured that the expectation $\mathbb{E}(\rho_t)$ converges to a point \tilde{x} in the proximity of the global minimizer x^* . On the other hand, in our works [CBO-I; CBO-II] we put forward a holistic analytical approach by investigating the time-evolution of the Wasserstein-2 distance $W_2^2(\rho_t, \delta_{x^*})$ and proving that it can be made as small as desired for suitable choices of the parameters and if the dynamics runs sufficiently long.

A Variance-Based Convergence Analysis¹²

Successful applications of CBO methods underlie the premise that the particle density ρ_t converges to a Dirac delta $\delta_{\tilde{x}}$ for some point \tilde{x} close to x^* . The analyses in [Car+18; Car+21] prove this under certain assumptions by first showing that ρ_t converges to a Dirac delta around some $\tilde{x} \in \mathbb{R}^d$ and then concluding $\tilde{x} \approx x^*$ in a subsequent step.

For the first step of the analysis, the variance $\text{Var}(\rho_t) := \frac{1}{2} \int \|x - \mathbb{E}(\rho_t)\|_2^2 d\rho_t(x)$, where $\mathbb{E}(\rho_t) := \int x d\rho_t(x)$, is shown to decay exponentially fast in time under a well-preparedness assumption about the initial condition ρ_0 . More precisely, in [Car+21, Appendix A] the authors use Itô's lemma to derive for the time-evolution of $\text{Var}(\rho_t)$ the expression

$$\frac{d}{dt} \text{Var}(\rho_t) = -(2\lambda - \sigma^2) \text{Var}(\rho_t) + \frac{\sigma^2}{2} \|\mathbb{E}(\rho_t) - x_\alpha^\mathcal{E}(\rho_t)\|_2^2. \quad (3.11)$$

For parameter choices $2\lambda > \sigma^2$, the first term in (3.11) is negative and one could almost apply Grönwall's inequality to obtain the asserted exponential decay of $\text{Var}(\rho_t)$. However, the second term in (3.11) is source of concern and the main difficulty is to control the distance $\|\mathbb{E}(\rho_t) - x_\alpha^\mathcal{E}(\rho_t)\|_2$ between the mean and the consensus point, i.e., the weighted mean. For $\alpha \rightarrow 0$ the weight function $\omega_\alpha^\mathcal{E}(\bullet) = \exp(-\alpha\mathcal{E}(\bullet))$ associated with $x_\alpha^\mathcal{E}(\rho_t)$ converges to 1 pointwise and consequently $x_\alpha^\mathcal{E}(\rho_t) \rightarrow \mathbb{E}(\rho_t)$. However, the second proof step, explained below, reveals that the crucial regime is $\alpha \gg 1$. In this case $x_\alpha^\mathcal{E}(\rho_t)$ can be arbitrarily far from $\mathbb{E}(\rho_t)$ if we do not dispose of additional knowledge about the probability measure ρ_t . To restrict the set of probability measures ρ_t that need to be considered when bounding $\|\mathbb{E}(\rho_t) - x_\alpha^\mathcal{E}(\rho_t)\|_2$, the authors of [Car+18; Car+21] compromise to assume that the initial distribution ρ_0 satisfies the well-preparedness assumptions

$$\mu_1 := 2\lambda - \sigma^2 - \frac{2\sigma^2 e^{-\alpha\mathcal{E}}}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\rho_0)}} > 0 \quad \text{and} \quad \mu_2 := (2\lambda + \sigma^2) \frac{\alpha e^{-2\alpha\mathcal{E}} C_5}{\mu_1 \|\omega_\alpha^\mathcal{E}\|_{L^1(\rho_0)}^2} \text{Var}(\rho_0) \leq \frac{3}{8}, \quad (3.12)$$

where $C_5 = \max(\|\max_i |\partial_{ii}\mathcal{E}|\|_\infty, \|\sigma(\nabla^2\mathcal{E})\|_\infty)$. Since ρ_t evolves from ρ_0 according to the Fokker-Planck equation (3.3), these conditions restrict ρ_t and allow for bounding $\|\mathbb{E}(\rho_t) - x_\alpha^\mathcal{E}(\rho_t)\|_2$ by a suitable multiple of $\text{Var}(\rho_t)$. The exponential decay of $\text{Var}(\rho_t)$ then follows from (3.11) after applying Grönwall's inequality, see [Car+21, Theorem 3.2]. Furthermore, the conditions in (3.12) also allow for proving convergence of $\mathbb{E}(\rho_t)$ to a stationary point $\tilde{x} \in \mathbb{R}^d$, see [Car+21, Theorem 3.2].

Given convergence to a Dirac at \tilde{x} , in a second step it is shown $\mathcal{E}(\tilde{x}) \approx \mathcal{E}(x^*)$. In order to prove this approximation, one first deduces that for any $\varepsilon > 0$, there exists

¹²In this section, we follow [CBO-I, Section 2.1] adapted to the setting of anisotropic noise [CBO-II].

$\alpha \gg 1$ such that for all $t \geq 0$ it holds $-\frac{1}{\alpha} \log(\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_t)}) \leq -\frac{1}{\alpha} \log(\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_0)}) + \frac{\varepsilon}{2}$. This involves deriving a lower bound for the evolution $\frac{d}{dt} \|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_t)}^2$ for sufficiently large $\alpha > 0$ as done in [Car+21, Lemma A.1], which is then combined with the formerly proven exponentially decaying variance, see [Car+21, Proof of Theorem 3.2]. Then, by recognizing that the Laplace principle (2.4) implies the existence of some $\alpha \gg 1$ with

$$-\frac{1}{\alpha} \log(\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_0)}) - \mathcal{E} < \frac{\varepsilon}{2}, \quad (3.13)$$

and by establishing the convergence $\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_t)} \rightarrow \exp(-\alpha\mathcal{E}(\tilde{x}))$ as $t \rightarrow \infty$, one obtains the desired result $\mathcal{E}(\tilde{x}) - \mathcal{E} < \varepsilon$ in the limit $t \rightarrow \infty$, see [Car+21, Proof of Theorem 3.2]. The gap $\mathcal{E}(\tilde{x}) - \mathcal{E}$ can be tightened by increasing α , but it is impossible to establish an explicit relation $\alpha = \alpha(\varepsilon)$ due to the use of the asymptotic Laplace principle (2.4).

This proof sketch unveils a tension on the role of the parameter α . Namely, the second step requires large $\alpha = \alpha(\varepsilon)$ to achieve $\mathcal{E}(\tilde{x}) - \mathcal{E} < \varepsilon$. In fact, $\alpha(\varepsilon)$ may grow uncontrollably as we decrease the accuracy ε . The first step, however, requires the well-preparedness conditions in (3.12) to hold, which, in the most optimistic case, where $\sigma = 0$, imply

$$\text{Var}(\rho_0) \leq \frac{3\mu_1}{8C_5\alpha} \left(\int \exp(-\alpha(\mathcal{E}(x) - \mathcal{E})) d\rho_0(x) \right)^2. \quad (3.14)$$

Therefore, ρ_0 needs to be increasingly concentrated as α increases, and should ideally be supported on sets where $\mathcal{E}(x) \approx \mathcal{E}$. Designing such distribution ρ_0 in practice seems impossible in the absence of a good initial guess for x^* . In particular, we cannot expect (3.14) to hold for generic choices such as a uniform distribution on a compact set.

Let us conclude this review by remarking that the works [HJK20; HJK21] conduct a similarly flavored analysis for the discrete-time microscopic system (2.2), with some differences in the details. They first show an exponentially decaying variance under mild assumptions about λ and σ , but provided that the same Brownian motion is used for all agents, i.e., $(B_k^i)_{k=1,\dots,K} = (B_k)_{k=1,\dots,K}$ for all $i = 1, \dots, N$. Such a choice leads to a considerably less explorative dynamics, but it simplifies the consensus formation analysis. For proving $\mathcal{E}(\tilde{x}) \approx \mathcal{E}$, however, the authors again require an initial configuration ρ_0 that satisfies a technical concentration condition like (3.13), see for instance [HJK21, Remark 3.1].

A Wasserstein Distance-Based Convergence Analysis¹³

The variance-based analysis approach described in the previous section has two drawbacks. Firstly, the analysis seems motivated by the technical expectation that the variance must vanish if the CBO method reaches any consensus, which does not shed light on the internal CBO mechanisms that lead to a successful minimization of the objective \mathcal{E} . Secondly, well-preparedness conditions such as (3.14), which severely restrict the initial configuration ρ_0 , are undesirable because they suggest that CBO methods are only successful if we have an informative initial guess about the location of the global

¹³In this section, we follow [CBO-I, Section 2.2] adapted to the setting of anisotropic noise [CBO-II].

minimizer x^* , which gives the results a certain locality flavor. To remedy these issues, let us now sketch and motivate the analytical framework proposed in our works [CBO-I; CBO-II].

By averaging out the randomness associated with different realizations of Brownian motion paths, the macroscopic continuous-time dynamics (3.2) becomes

$$\begin{aligned} \frac{d}{dt} \mathbb{E} [\bar{X}_t | \bar{X}_0] &= -\lambda \mathbb{E} \left[\left(\bar{X}_t - x_\alpha^\mathcal{E}(\rho_t) \right) \middle| \bar{X}_0 \right] \\ &= -\lambda \mathbb{E} \left[\left(\bar{X}_t - x^* \right) \middle| \bar{X}_0 \right] + \lambda \left(x_\alpha^\mathcal{E}(\rho_t) - x^* \right). \end{aligned} \quad (3.15)$$

Under the assumption that the objective \mathcal{E} is locally Lipschitz continuous and satisfies a coercivity condition with parameters $\eta > 0$ and $\nu \in (0, \infty)$ of the form

$$\|x - x^*\|_\infty \leq \frac{1}{\eta} (\mathcal{E}(x) - \mathcal{E}(x^*))^\nu = \frac{1}{\eta} (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu, \quad \text{for all } x \in \mathbb{R}^d, \quad (3.16)$$

the last term in (3.15) can be made arbitrarily small for a sufficiently large parameter α i.e., $x_\alpha^\mathcal{E}(\rho_t) \approx x^*$, by a quantitative version of the Laplace principle, see Proposition 3.12. In this case, the average dynamics of \bar{X}_t is well-approximated by

$$\frac{d}{dt} \mathbb{E} [\bar{X}_t | \bar{X}_0] \approx -\lambda \mathbb{E} \left[\left(\bar{X}_t - x^* \right) \middle| \bar{X}_0 \right], \quad (3.17)$$

which corresponds to the gradient flow of $x \mapsto \|x - x^*\|_2^2$ with rate 2λ . In other words, each individual agent essentially performs in the mean-field limit a gradient descent of $x \mapsto \|x - x^*\|_2^2$ on average over all realizations of Brownian motion paths. Figure 3.1b visualizes this phenomenon for three isolated agents on the Rastrigin function in two dimensions, which is depicted in Figure 3.1a.

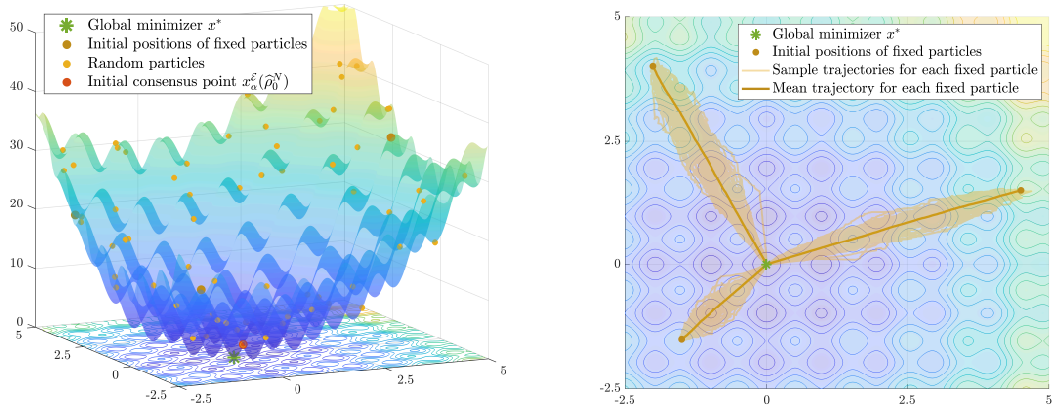
Inspired by this observation, our proof strategy is to show that CBO methods minimize the functional $\mathcal{V} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}_{\geq 0}$, given by

$$\mathcal{V}(\rho_t) := \frac{1}{2} \int \|x - x^*\|_2^2 d\rho_t(x). \quad (3.18)$$

Note that this functional essentially coincides with the Wasserstein-2 distance to a Dirac delta δ_{x^*} located at the global minimizer x^* of \mathcal{E} . In formulas, $W_2^2(\rho_t, \delta_{x^*}) = 2\mathcal{V}(\rho_t)$. Therefore, $\mathcal{V}(\rho_t) \rightarrow 0$ implies that ρ_t converges weakly to δ_{x^*} , see [AGS08, Chapter 7], giving consensus formation as a byproduct. The latter can be seen when noticing that $\mathcal{V}(\rho_t)$ majorizes $\text{Var}(\rho_t)$ because $z \mapsto \frac{1}{2} \int \|x - z\|_2^2 d\rho_t(x)$ is minimized by the expectation $\mathbb{E}(\rho_t)$.

This approach does not suffer a tension on the parameter α like the variance-based analysis framework from the previous section. For the functional $\mathcal{V}(\rho_t)$ we establish in Lemma 3.10 an evolution inequality of the form

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\leq - (2\lambda - \sigma^2) \mathcal{V}(\rho_t) + \sqrt{2}(\lambda + \sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2 \\ &\quad + \frac{\sigma^2}{2} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2^2, \end{aligned} \quad (3.19)$$



(a) The highly nonconvex Rastrigin function in two dimensions as objective function \mathcal{E} and an exemplary initialization for one run of the experiment.

(b) CBO performs a canonical convexification in the mean-field limit.

Individual agents follow, on average, the gradient flow of the map $x \mapsto \|x - x^*\|_2^2$, which is independent of the underlying energy landscape of \mathcal{E} .

Figure 3.1: An illustration of the internal mechanisms of CBO. We perform 100 runs of the CBO algorithm (2.2), with parameters $\Delta t = 0.01$, $\alpha = 10^{15}$, $\lambda = 1$ and $\sigma = 0.1$, and $N = 32000$ agents initialized according to $\rho_0 = \mathcal{N}((8, 8), 20)$. In addition, we add three individual agents with starting locations $(-2, 4)$, $(-1.5, -1.5)$ and $(4.5, 1.5)$ to the set of agents in each run as shown in (a), and depict each of their 100 trajectories as well as their mean trajectory in yellow color in (b). With the (mean) trajectories being rather straight lines, we observe that the individual agents take a straight path from their initial positions to the global minimizer x^* (green star), in particular, disregard the local landscape of the objective function \mathcal{E} . The trajectories of the individual agents become more concentrated as the overall number of agents N grows and for smaller values of the diffusion parameter σ .

where it remains to control the term $\|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2$. However, in comparison to bounding $\|x_\alpha^\mathcal{E}(\rho_t) - \mathbb{E}(\rho_t)\|_2$ as was necessary for (3.11) in the variance-based analysis, this is a much easier and more natural task. Under the inverse continuity condition (3.16), the Laplace principle (2.4) asserts $\|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2 \rightarrow 0$ as $\alpha \rightarrow \infty$. Quantitatively, for an arbitrary probability measure ϱ , we can even establish

$$\|x_\alpha^\mathcal{E}(\varrho) - x^*\|_2 \leq \frac{\sqrt{d}(2Lr)^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha Lr)}{\varrho(B_r^\infty(x^*))} \int \|x - x^*\|_2 d\varrho(x) \quad (3.20)$$

as follows from Proposition 3.12 in the setting of (3.16) and assuming that \mathcal{E} is L -Lipschitz in a ball of radius $r > 0$ around x^* . A similar result holds under the more general version of the inverse continuity condition A2 in Assumption 3.5. Therefore, choosing a small radius $r > 0$ and large $\alpha > 0$ accordingly allows for controlling $\|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2$ and we do not suffer any of the drawbacks raised in the previous section related to large choices of α . In particular, analogously to the more detailed proof for the isotropic noise

case in [CBO-I, Section 4.4], which holds mutatis mutandis in the anisotropic setting as we describe after [Theorem 3.6](#), we can find, for a given accuracy $\varepsilon \in (0, \mathcal{V}(\rho_0))$, a threshold $\alpha_0(\varepsilon, \vartheta, \mathcal{E})$ such that for $\alpha > \alpha_0$ it holds

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -(1 - \vartheta)(2\lambda - \sigma^2) \mathcal{V}(\rho_t) \quad (3.21)$$

for all $t \in (0, T_\alpha)$, where T_α denotes the time when the functional \mathcal{V} first achieves the desired accuracy ε , i.e., $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$. T_α may depend on α . This requires that the probability mass w.r.t. the measure ρ_t of arbitrarily small ℓ^∞ -balls $B_r^\infty(x^*)$ around x^* does not vanish. Provided that $x^* \in \text{supp}(\rho_0)$, [Proposition 3.13](#) ensures the latter for a finite time horizon by devising an estimate of the form $\rho_t(B_r^\infty(x^*)) \gtrsim \rho_0(B_{r/2}^\infty(x^*)) \exp(-qt)$, i.e., the initial mass $\rho_0(B_{r/2}^\infty(x^*)) > 0$ can decay at most at an exponential rate for any $r > 0$, but remains strictly positive in any finite time window $[0, T_\alpha]$. A key requirement to this result is an active diffusion term, i.e., $\sigma > 0$, which counteracts the deterministic movement of the drift term by inducing randomness. On the other hand, with similar arguments and in analogy to (3.19) and (3.21), we can derive with [Lemma 3.11](#) an evolution inequality for the functional $\mathcal{V}(\rho_t)$ of the form

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\geq -(2\lambda - \sigma^2) \mathcal{V}(\rho_t) - \sqrt{2}(\lambda + \sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2 \\ &\geq -(1 + \vartheta/2)(2\lambda - \sigma^2) \mathcal{V}(\rho_t) \end{aligned} \quad (3.22)$$

for all $t \in (0, T_\alpha)$. Grönwall's inequality now implies for all $t \in [0, T_\alpha]$ the upper and lower bound

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0) \exp\left(-(1 - \vartheta)(2\lambda - \sigma^2)t\right), \quad (3.23)$$

$$\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0) \exp\left(-(1 + \vartheta/2)(2\lambda - \sigma^2)t\right), \quad (3.24)$$

i.e., $\mathcal{V}(\rho_t)$ decays at least exponentially fast (with rate $(1 - \vartheta)(2\lambda - \sigma^2)$), and at most exponentially fast (with rate $(1 + \vartheta/2)(2\lambda - \sigma^2)$). With the true decay behavior of $\mathcal{V}(\rho_t)$ being sandwiched between these decay rates and after recalling the definition of T_α , we can infer that

$$\frac{1 - \vartheta}{(1 + \vartheta/2)} T^* = \frac{1}{(1 + \vartheta/2)(2\lambda - \sigma^2)} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) \leq T_\alpha = T^*, \quad (3.25)$$

where $T^* := \frac{1}{(1 - \vartheta)(2\lambda - \sigma^2)} \log(\mathcal{V}(\rho_0)/\varepsilon)$ is as in (3.28) below.

Following this intuition and the associated proof sketch, let us now present the main result about global convergence of anisotropic CBO in mean-field law for objective functions satisfying the following.

Assumption 3.5. Throughout we are interested in objectives $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$, for which

- A1 there exists $x^* \in \mathbb{R}^d$ such that $\mathcal{E}(x^*) = \inf_{x \in \mathbb{R}^d} \mathcal{E}(x) =: \underline{\mathcal{E}}$,

A2 there exist $\mathcal{E}_\infty, R_0, \eta > 0$, and $\nu \in (0, \infty)$ such that

$$\|x - x^*\|_\infty \leq (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu / \eta, \quad \text{for all } x \in B_{R_0}^\infty(x^*), \quad (3.26)$$

$$\mathcal{E}(x) - \underline{\mathcal{E}} > \mathcal{E}_\infty, \quad \text{for all } x \in (B_{R_0}^\infty(x^*))^c. \quad (3.27)$$

A1 just states that the continuous objective \mathcal{E} attains its infimum $\underline{\mathcal{E}}$ at some $x^* \in \mathbb{R}^d$. A2 should be interpreted as a tractability condition of the landscape of \mathcal{E} around x^* and in the farfield. The first part, Equation (3.26), describes the local coercivity of \mathcal{E} , which implies that there is a unique minimizer x^* on $B_{R_0}^\infty(x^*)$ and that \mathcal{E} grows like $x \mapsto \|x - x^*\|_\infty^{1/\nu}$. This condition is also known as the inverse continuity condition from [For+21], as the quadratic growth condition in the case $\nu = 1/2$ from [Ani00; NNG19], as the Hölderian error bound condition in the case $\nu \in (0, 1]$ [Bol+17], or as the $1/\nu$ -conditioning property from [GRV23, Definition 3.1]. In [NNG19, Theorem 4] and [KNS16, Theorem 2] many equivalent or stronger conditions are identified to imply Equation (3.26) globally on \mathbb{R}^d . Furthermore, in [XLY17; For+21], (3.26) is shown to hold globally for objectives related to various machine learning problems. The second part of A2, Equation (3.27), describes the behavior of \mathcal{E} in the farfield and prevents $\mathcal{E}(x) \approx \underline{\mathcal{E}}$ for some $x \in \mathbb{R}^d$ far away from x^* . We introduce it for the purpose of covering functions that tend to a constant just above \mathcal{E}_∞ as $\|x\|_\infty \rightarrow \infty$, because such functions do not satisfy the growth condition (3.26) globally. Together with (3.26) it implies the uniqueness of the global minimizer x^* on the whole \mathbb{R}^d . However, whenever (3.26) holds globally, we take $R_0 = \infty$, i.e., $B_{R_0}^\infty(x^*) = \mathbb{R}^d$ and (3.27) is void.

Under these assumptions, we have the following statement, which is an improvement of [CBO-II, Theorem 2].

Theorem 3.6 (CBO converges globally in mean-field law, cf. [CBO-I, Theorem 12]). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ be such that $x^* \in \text{supp}(\rho_0)$. Define $\mathcal{V}(\rho_t)$ as given in (3.18). Provided that $\mathcal{V}(\rho_0) > 0$, fix any $\varepsilon \in (0, \mathcal{V}(\rho_0))$ and $\vartheta \in (0, 1)$, choose parameters $\lambda, \sigma > 0$ with $2\lambda > \sigma^2$, and define the time horizon

$$T^* := \frac{1}{(1 - \vartheta)(2\lambda - \sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right). \quad (3.28)$$

Then there exists $\alpha_0 > 0$, depending (among problem dependent quantities) on ε and ϑ , such that for all $\alpha > \alpha_0$, if $\rho \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (3.3) on the time interval $[0, T^*]$ with initial condition ρ_0 , we have

$$\mathcal{V}(\rho_T) = \varepsilon \quad \text{with} \quad T \in \left[\frac{1 - \vartheta}{(1 + \vartheta/2)} T^*, T^* \right]. \quad (3.29)$$

Furthermore, on the time interval $[0, T]$, $\mathcal{V}(\rho_t)$ decays exponentially fast. More precisely, for all $t \in [0, T]$, it holds

$$W_2^2(\rho_t, \delta_{x^*}) = 2\mathcal{V}(\rho_t) \leq 2\mathcal{V}(\rho_0) \exp \left(-(1 - \vartheta)(2\lambda - \sigma^2)t \right) \quad (3.30)$$

as well as

$$W_2^2(\rho_t, \delta_{x^*}) = 2\mathcal{V}(\rho_t) \geq 2\mathcal{V}(\rho_0) \exp\left(- (1 + \vartheta/2)(2\lambda - \sigma^2)t\right). \quad (3.31)$$

Before presenting the proof of [Theorem 3.6](#), let us provide some remarks about different facets of the result. Afterwards, we give some auxiliary results for the proof, which may be of independent interest.

Remark 3.7. The statement of [Theorem 3.6](#) is valid for any $\rho \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d))$ weakly solving the Fokker-Planck equation [\(3.3\)](#) with initial datum ρ_0 . Sufficient conditions for the existence of such ρ are provided by the assumptions [W1–W2](#) of [Theorem 3.4](#).

Remark 3.8 (Convergence rate $(2\lambda - \sigma^2)$, cf. [\[CBO-I, Section 3.2\]](#)). Lower and upper bounds on the rate of convergence of $\mathcal{V}(\rho_t)$ are $(1 - \vartheta)(2\lambda - \sigma^2)$ and $(1 + \vartheta/2)(2\lambda - \sigma^2)$, see [\(3.30\)](#) and [\(3.31\)](#), respectively, which can be made arbitrarily close to the numerically observed rate $(2\lambda - \sigma^2)$, see, e.g., [\[CBO-II, Figure 1\(b\)\]](#), at the cost of taking $\alpha \rightarrow \infty$ to allow for $\vartheta \rightarrow 0$. The condition $2\lambda > \sigma^2$ is necessary, both in theory and practice, to avoid overwhelming the dynamics by the random exploration term.

Remark 3.9 (Initial configuration ρ_0 , cf. [\[CBO-I, Section 3.2\]](#)). The assumption $x^* \in \text{supp}(\rho_0)$ about the initial configuration ρ_0 is not really a restriction, as it would anyhow hold immediately for ρ_t for any $t > 0$ in view of the diffusive character of the mean-field dynamics [\(3.3\)](#), see [Remark 3.14](#). Additionally, as we clarify in [Section 3.1.2](#), this condition does neither mean nor require that, for finite particle approximations, some particle needs to be in the vicinity of the minimizer x^* at time $t = 0$. It is actually sufficient that the empirical measure $\hat{\rho}_t^N$ weakly approximates the law ρ_t uniformly in time. We rigorously explain this mechanism in [Section 3.1.3](#).

A comment on the pivotal role of the parameter α , is postponed to [Remark 3.20](#) in [Section 3.1.3](#).

Lemma 3.10 (Time-evolution of \mathcal{V} , [\[CBO-II, Lemma 1\]](#)). Let $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ and fix $\alpha, \lambda, \sigma > 0$. Moreover, let $T > 0$ and let $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ be a weak solution to the Fokker-Planck equation [\(3.3\)](#). Then the functional $\mathcal{V}(\rho_t)$ satisfies

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\leq - (2\lambda - \sigma^2) \mathcal{V}(\rho_t) + \sqrt{2} (\lambda + \sigma^2) \sqrt{\mathcal{V}(\rho_t)} \left\| x_\alpha^\mathcal{E}(\rho_t) - x^* \right\|_2 \\ &\quad + \frac{\sigma^2}{2} \left\| x_\alpha^\mathcal{E}(\rho_t) - x^* \right\|_2^2. \end{aligned} \quad (3.32)$$

The proof of [Lemma 3.10](#) is presented in [\[CBO-II, Lemma 1\]](#), see also [\[CBO-I, Lemma 17\]](#) for a more detailed proof in the isotropic case.

Lemma 3.11 (Time-evolution of \mathcal{V}). Under the assumptions of Lemma 3.10, the functional $\mathcal{V}(\rho_t)$ satisfies

$$\frac{d}{dt}\mathcal{V}(\rho_t) \geq -\left(2\lambda - \sigma^2\right)\mathcal{V}(\rho_t) - \sqrt{2}\left(\lambda + \sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2. \quad (3.33)$$

The proof of Lemma 3.11 follows analogously to the one of [CBO-I, Lemma 18] in the isotropic case.

In order to apply Grönwall's inequality to (3.32) and (3.33), we need to control the term $\|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2$, which is the task of the following result.

Proposition 3.12 (Quantitative Laplace principle, [CBO-II, Proposition 1]).

Let $\underline{\mathcal{E}} = 0$, $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$ we define $\mathcal{E}_r := \sup_{x \in B_r^\infty(x^*)} \mathcal{E}(x)$. Then, under the inverse continuity property A2, for any $r \in (0, R_0]$ and $q > 0$ such that $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, we have

$$\|x_\alpha^\mathcal{E}(\varrho) - x^*\|_2 \leq \frac{\sqrt{d}(q + \mathcal{E}_r)^\nu}{\eta} + \frac{\sqrt{d}\exp(-\alpha q)}{\varrho(B_r^\infty(x^*))} \int \|x - x^*\|_2 d\varrho(x). \quad (3.34)$$

The proof of Proposition 3.12 is presented in [CBO-II, Section 3.3].

To apply Proposition 3.12 in the proof of Theorem 3.6, we require a lower bound for the probability mass of $\rho_t(B_r^\infty(x^*))$, where $r > 0$ is a small radius. This is achieved by defining a mollifier $\phi_r : \mathbb{R}^d \rightarrow \mathbb{R}$ according to

$$\phi_r(x) := \begin{cases} \prod_{k=1}^d \exp\left(1 - \frac{r^2}{r^2 - (x - x^*)_k^2}\right), & \text{if } \|x - x^*\|_\infty < r, \\ 0, & \text{else.} \end{cases} \quad (3.35)$$

Since $\rho_t(B_r^\infty(x^*)) \geq \int \phi_r(x) d\rho_t(x)$, the desired lower bound can be obtained by studying the evolution of the right-hand side.

Proposition 3.13 (A lower bound for the probability mass around x^* , [CBO-II, Proposition 2]).

Let $T > 0$, $r > 0$, and fix parameters $\alpha, \lambda, \sigma > 0$. Assume $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ weakly solves the Fokker-Planck equation (3.3) with initial condition $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ and for $t \in [0, T]$. Then, for all $t \in [0, T]$ we have

$$\begin{aligned} \rho_t(B_r^\infty(x^*)) &= \rho_t\left(\left\{x \in \mathbb{R}^d : \|x - x^*\|_\infty \leq r\right\}\right) \\ &\geq \left(\int \phi_r(x) d\rho_0(x)\right) \exp(-qt) \end{aligned} \quad (3.36)$$

with

$$q := 2d \max \left\{ \frac{\lambda(cr + B\sqrt{c})}{(1-c)^2 r} + \frac{\sigma^2(cr^2 + B^2)(2c+1)}{(1-c)^4 r^2}, \frac{2\lambda^2}{(2c-1)\sigma^2} \right\}, \quad (3.37)$$

for any $B < \infty$ with $\sup_{t \in [0, T]} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_\infty \leq B$ and for any $c \in (1/2, 1)$ satisfying $(1-c)^2 \leq (2c-1)c$.

The proof of [Proposition 3.13](#) is presented in [[CBO-II](#), Section 3.4].

Remark 3.14 (Necessity of the diffusion σ , cf. [[CBO-I](#), Remark 23]). [Proposition 3.13](#) demonstrates the crucial role of positive σ in the stochastic terms in [\(2.2\)](#) and [\(3.1\)](#), or the diffusion in the macroscopic models [\(3.2\)](#) and [\(3.3\)](#) for our analysis. $\sigma > 0$ is required in [\(3.37\)](#) to warrant a finite decay rate $q < \infty$. Thanks to [\(3.36\)](#), this, in turn, guarantees mass around the minimizer x^* , ensuring the applicability of the quantitative Laplace principle in [Proposition 3.12](#) at every point in time $t \in [0, T]$. Intuitively, we can understand the measure ρ as having a deterministic component, which evolves according to the drift term in the Fokker-Planck equation [\(3.3\)](#) and whose associated mass may leave $B_r^\infty(x^*)$ in finite time, convolved with an exponentially decaying kernel from the diffusion term. This convolution ensures that the mass leaves at most exponentially fast, leading to the lower bound. The statement does not hold in general for the case $\sigma = 0$.

We now have the necessary technical tools to provide the proof of [Theorem 3.6](#).

Proof of [Theorem 3.6](#). If $\mathcal{V}(\rho_0) = 0$, there is nothing to be shown since in this case $\rho_0 = \delta_{x^*}$. Thus, let $\mathcal{V}(\rho_0) > 0$ in what follows.

W.l.o.g. we may assume $\underline{\mathcal{E}} = 0$. Let us first choose the parameter α such that

$$\alpha > \alpha_0 := \frac{1}{q_\varepsilon} \left(\log \left(\frac{2^{d+1} \sqrt{2d\mathcal{V}(\rho_0)}}{c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon}} \right) + \frac{q}{(1-\vartheta)(2\lambda - \sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) - \log \rho_0(B_{r_\varepsilon/2}^\infty(x^*)) \right), \quad (3.38)$$

where we introduce the definitions

$$c(\vartheta, \lambda, \sigma) := \min \left\{ \frac{\vartheta}{2} \frac{(2\lambda - \sigma^2)}{\sqrt{2}(\lambda + \sigma^2)}, \sqrt{\vartheta} \frac{(2\lambda - \sigma^2)}{\sigma^2} \right\} \quad (3.39)$$

as well as

$$q_\varepsilon := \frac{1}{2} \min \left\{ \left(\eta \frac{c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon}}{2\sqrt{d}} \right)^{1/\nu}, \mathcal{E}_\infty \right\} \quad \text{and} \quad r_\varepsilon := \max_{s \in [0, R_0]} \left\{ \max_{x \in B_s^\infty(x^*)} \mathcal{E}(x) \leq q_\varepsilon \right\}. \quad (3.40)$$

Moreover, q is as defined in [\(3.37\)](#) with $B = c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_0)}$ and with $r = r_\varepsilon$. We remark that, by construction, $q_\varepsilon > 0$ and $r_\varepsilon \leq R_0$. Furthermore, recalling the notation $\mathcal{E}_r = \sup_{x \in B_r^\infty(x^*)} \mathcal{E}(x)$ from [Proposition 3.12](#), we have $q_\varepsilon + \mathcal{E}_{r_\varepsilon} \leq 2q_\varepsilon \leq \mathcal{E}_\infty$ as a consequence of the definition of r_ε . Since $q_\varepsilon > 0$, the continuity of \mathcal{E} ensures that there exists $s_{q_\varepsilon} > 0$ such that $\mathcal{E}(x) \leq q_\varepsilon$ for all $x \in B_{s_{q_\varepsilon}}^\infty(x^*)$, thus yielding also $r_\varepsilon > 0$.

Let us now define the time horizon $T_\alpha \geq 0$, which may depend on α , by

$$T_\alpha := \sup \left\{ t \geq 0 : \mathcal{V}(\rho_t) > \varepsilon \text{ and } \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2 < C(t) \text{ for all } t' \in [0, t] \right\} \quad (3.41)$$

with $C(t) := c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_t)}$. Notice for later use that $C(0) = B$.

Our aim now is to show $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in [\frac{1-\vartheta}{(1+\vartheta/2)} T^*, T^*]$ and that we have at least exponential decay of $\mathcal{V}(\rho_t)$ until time T_α , i.e., until accuracy ε is reached.

First, however, we ensure that $T_\alpha > 0$. With the mapping $t \mapsto \mathcal{V}(\rho_t)$ being continuous as a consequence of the regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ established in [CBO-II, Theorem 1] and $t \mapsto \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2$ being continuous due to [Car+18, Lemma 3.2] and $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$, $T_\alpha > 0$ follows from the definition, since $\mathcal{V}(\rho_0) > \varepsilon$ and $\|x_\alpha^\mathcal{E}(\rho_0) - x^*\|_2 < C(0)$. While the former is immediate by assumption, applying Proposition 3.12 with q_ε and r_ε gives the latter since

$$\begin{aligned} \|x_\alpha^\mathcal{E}(\rho_0) - x^*\|_2 &\leq \frac{\sqrt{d}(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_0(B_{r_\varepsilon}^\infty(x^*))} \int \|x - x^*\|_2 d\rho_0(x) \\ &\leq \frac{c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon}}{2} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_0(B_{r_\varepsilon}^\infty(x^*))} \sqrt{2\mathcal{V}(\rho_0)} \\ &\leq c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon} \\ &< c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_0)} = C(0), \end{aligned} \quad (3.42)$$

where the inequality in the next-to-last line holds by the choice of α in (3.38).

Next, we show that the functional $\mathcal{V}(\rho_t)$ decays essentially exponentially fast in time. More precisely, we prove that, up to time T_α , $\mathcal{V}(\rho_t)$ decays

- (i) at least exponentially fast (with rate $(1 - \vartheta)(2\lambda - \sigma^2)$), and
- (ii) at most exponentially fast (with rate $(1 + \vartheta/2)(2\lambda - \sigma^2)$).

To obtain (i), recall that Lemma 3.10 provides an upper bound on $\frac{d}{dt}\mathcal{V}(\rho_t)$ given by

$$\begin{aligned} \frac{d}{dt}\mathcal{V}(\rho_t) &\leq -(2\lambda - \sigma^2)\mathcal{V}(\rho_t) + \sqrt{2}(\lambda + \sigma^2)\sqrt{\mathcal{V}(\rho_t)} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2 \\ &\quad + \frac{\sigma^2}{2} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2^2. \end{aligned} \quad (3.43)$$

Combining this with the definition of T_α in (3.41) we have by construction

$$\frac{d}{dt}\mathcal{V}(\rho_t) \leq -(1 - \vartheta)(2\lambda - \sigma^2)\mathcal{V}(\rho_t), \quad \text{for all } t \in (0, T_\alpha). \quad (3.44)$$

For (ii), on the other hand, by using Lemma 3.11 we can derive a lower bound on $\frac{d}{dt}\mathcal{V}(\rho_t)$ of the form

$$\begin{aligned} \frac{d}{dt}\mathcal{V}(\rho_t) &\geq -(2\lambda - \sigma^2)\mathcal{V}(\rho_t) - \sqrt{2}(\lambda + \sigma^2)\sqrt{\mathcal{V}(\rho_t)} \|x_\alpha^\mathcal{E}(\rho_t) - x^*\|_2 \\ &\geq -(1 + \vartheta/2)(2\lambda - \sigma^2)\mathcal{V}(\rho_t), \quad \text{for all } t \in (0, T_\alpha), \end{aligned}$$

where the second inequality again exploits the definition of T_α . Grönwall's inequality now implies for all $t \in [0, T_\alpha]$ the upper and lower bound

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0) \exp\left(-(1 - \vartheta)(2\lambda - \sigma^2)t\right), \quad (3.45)$$

$$\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0) \exp\left(-(1 + \vartheta/2)(2\lambda - \sigma^2)t\right), \quad (3.46)$$

thereby proving (i) and (ii). We further note that the definition of T_α in (3.41) together with the definition of $C(t)$ and (3.45) permits to control

$$\max_{t \in [0, T_\alpha]} \|x_\alpha^\varepsilon(\rho_t) - x^*\|_2 \leq \max_{t \in [0, T_\alpha]} C(t) \leq C(0). \quad (3.47)$$

To conclude, it remains to prove that $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in [\frac{1-\vartheta}{(1+\vartheta/2)} T^*, T^*]$. For this we distinguish the following three cases.

Case $T_\alpha \geq T^*$: We can use the definition of T^* in (3.28) and the time-evolution bound of $\mathcal{V}(\rho_t)$ in (3.45) to conclude that $\mathcal{V}(\rho_{T^*}) \leq \varepsilon$. Hence, by definition of T_α in (3.41) together with the continuity of $\mathcal{V}(\rho_t)$, we find $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha = T^*$.

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) \leq \varepsilon$: By continuity of $\mathcal{V}(\rho_t)$, it holds for T_α as defined in (3.41), $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$. Thus, $\varepsilon = \mathcal{V}(\rho_{T_\alpha}) \geq \mathcal{V}(\rho_0) \exp(-(1+\vartheta/2)(2\lambda - \sigma^2)T_\alpha)$ by (3.46), which can be reordered as

$$\frac{1-\vartheta}{(1+\vartheta/2)} T^* = \frac{1}{(1+\vartheta/2)(2\lambda - \sigma^2)} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) \leq T_\alpha < T^*. \quad (3.48)$$

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$: We shall show that this case can never occur by verifying that $\|x_\alpha^\varepsilon(\rho_{T_\alpha}) - x^*\|_2 < C(T_\alpha)$ due to the choice of α in (3.38). In fact, fulfilling simultaneously both $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$ and $\|x_\alpha^\varepsilon(\rho_{T_\alpha}) - x^*\|_2 < C(T_\alpha)$ would contradict the definition of T_α in (3.41) itself. To this end, by applying again Proposition 3.12 with q_ε and r_ε , and recalling that $\varepsilon < \mathcal{V}(\rho_{T_\alpha})$, we get

$$\begin{aligned} \|x_\alpha^\varepsilon(\rho_{T_\alpha}) - x^*\|_2 &\leq \frac{\sqrt{d}(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{T_\alpha}(B_{r_\varepsilon}^\infty(x^*))} \int \|x - x^*\|_2 d\rho_{T_\alpha}(x) \\ &< \frac{c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{T_\alpha}(B_{r_\varepsilon}^\infty(x^*))} \sqrt{2\mathcal{V}(\rho_{T_\alpha})}. \end{aligned} \quad (3.49)$$

Since, thanks to (3.47), we have the bound $\max_{t \in [0, T_\alpha]} \|x_\alpha^\varepsilon(\rho_t) - x^*\|_2 \leq B$ for $B = C(0)$, which is in particular independent of α , Proposition 3.13 guarantees that there exists a $q > 0$ not depending on α (but depending on B and r_ε) with

$$\begin{aligned} \rho_{T_\alpha}(B_{r_\varepsilon}^\infty(x^*)) &\geq \left(\int \phi_{r_\varepsilon}(x) d\rho_0(x) \right) \exp(-qT_\alpha) \\ &\geq \frac{1}{2^d} \rho_0(B_{r_\varepsilon/2}^\infty(x^*)) \exp(-qT^*) > 0, \end{aligned}$$

where we used $x^* \in \text{supp}(\rho_0)$ for bounding the initial mass ρ_0 and the fact that ϕ_r (as defined in (3.35)) is bounded from below on $B_{r/2}^\infty(x^*)$ by $1/2^d$. With this we can continue the chain of inequalities in (3.49) to obtain

$$\begin{aligned} \|x_\alpha^\varepsilon(\rho_{T_\alpha}) - x^*\|_2 &< \frac{c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2^d \sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_0(B_{r_\varepsilon/2}^\infty(x^*)) \exp(-qT^*)} \sqrt{2\mathcal{V}(\rho_{T_\alpha})} \\ &\leq c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_{T_\alpha})} = C(T_\alpha), \end{aligned} \quad (3.50)$$

where the first inequality in the last line holds by the choice of α in (3.38). This establishes the desired contradiction, again as consequence of the continuity of the mappings $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \|x_\alpha^\varepsilon(\rho_t) - x^*\|_2$. \square

3.1.2. Mean-Field Approximation¹⁴

By investigating the CBO dynamics in the preceding [Section 3.1.1](#) from a mean-field perspective, we unveiled the surprising phenomenon that, for a rich class of objective functions \mathcal{E} , the mean-field CBO dynamics [\(3.3\)](#) performs a generic convexification of general nonconvex problems. This insight reveals that the hardness of any global optimization problem is necessarily encoded in the rate of the mean-field approximation as $N \rightarrow \infty$. In consideration of the central significance of such result with regards to the overall computational complexity of the numerical CBO scheme [\(2.2\)](#), this necessitates a quantitative result about the convergence of the interacting particle system [\(3.1\)](#) to the corresponding mean-field limit [\(3.2\)](#) and [\(3.3\)](#) in terms of the number of employed particles N .

As described intuitively but colloquially in the introduction of [Section 3.1](#), we expect that the individual agents of the interacting particle system [\(3.1\)](#) become statistically independent with their number N tending to infinity, due to their mutual influence on each other decreasing. More formally, as $N \rightarrow \infty$, we expect that the random empirical measure $\widehat{\rho}^N$ of the interacting particle system [\(3.1\)](#) converges in law to the deterministic distribution ρ of the mean-field dynamics [\(3.2\)](#) almost everywhere, i.e.,

$$\widehat{\rho}_t^N \rightharpoonup \rho_t \quad \text{as } N \rightarrow \infty \quad (3.51)$$

for almost every $t \geq 0$, see, e.g., [\[CD22a; CD22b\]](#) or [\[JW17\]](#) for extensive reviews on the topic of mean-field limits and approximation results for interacting particle systems.

The classical way to establish such mean-field approximation, proposed in the seminal work [\[McK67\]](#) and later extended and popularized by the author of [\[Szn91\]](#), is to prove propagation of chaos¹⁵ by means of the coupling method [\[CD22a, Section 4.1\]](#). By exploiting the SDE representation of the interacting particle system, thus requiring regularity of the microscopic system and typically well-posedness of the mean-field limit, this argument leads to quantitative mean-field approximation results w.r.t. the number of particles N . We elaborate on these ideas in what follows. Before that, however, let us mention that qualitative propagation of chaos statements can be obtained for wider classes of interacting particle systems through stochastic tightness and compactness methods [\[CD22a, Section 4.2\]](#), see, e.g., [\[Szn84; GM97\]](#) for two such exemplary settings. Again non-quantitative, but strong, yet abstract results going beyond but comprising propagation of chaos can be furthermore obtained with techniques related to large deviations [\[CD22a, Section 4.4\]](#) and [\[CD22b, Section 5.4\]](#).

¹⁴In this section, we follow [\[CBO-I, Remark 2\]](#) as well as [\[CBO-I, Section 3.3\]](#) adapted to the setting of anisotropic noise [\[CBO-II\]](#).

¹⁵In the sense of [\[Kac56; McK67\]](#), a distribution $\widehat{\varrho}^N$ that is invariant under perturbations is called ϱ -chaotic, if for any $k \leq N$ any k -marginal of $\widehat{\varrho}^N$ converges weakly to the product measure $\varrho^{\otimes k}$ as $N \rightarrow \infty$. An interacting particle system described through its random empirical measure $\widehat{\rho}^N$ is said to have the propagation of chaos property if chaos propagates through the system in time despite the interacting nature of the system. I.e., propagation of chaos holds if $\widehat{\rho}_0^N$ being ρ_0 -chaotic implies that, at any time $t > 0$, $\widehat{\rho}_t^N$ is ρ_t -chaotic for a suitable ρ . Colloquially speaking, if the particles are initialized i.i.d. at $t = 0$, thus being chaotic, they are not independent at any time point $t > 0$ anymore due to their interactions. However, in the large particle limit $N \rightarrow \infty$, the independence property is recovered.

Back to the coupling method, the most popular and at the same time simplest choice is the synchronous coupling [JW17, Section 3.1] and [CD22a, Section 4.1.2]. To the interacting particle dynamics, denoted by $(X^i)_{i=1,\dots,N}$ and described by the SDE system (3.1) endowed with initial data $(X_0^i)_{i=1,\dots,N}$ distributed i.i.d. according to $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ and driven by the Brownian motions $((B_t^i)_{t \geq 0})_{i=1,\dots,N}$, we couple the non-interacting system of N independent copies of the self-consistent SDE (3.2), i.e.,

$$d\bar{X}_t^i = -\lambda \left(\bar{X}_t^i - x_\alpha^\mathcal{E}(\rho_t) \right) dt + \sigma D \left(\bar{X}_t^i - x_\alpha^\mathcal{E}(\rho_t) \right) dB_t^i \quad (3.52)$$

with $\bar{X}_0^i = X_0^i$ for $i = 1, \dots, N$. It is straightforward to notice that, due to the independence of the Brownian motions $((B_t^i)_{t \geq 0})_{i=1,\dots,N}$, the processes $(\bar{X}^i)_{i=1,\dots,N}$ are indeed independent copies of (3.2) with $\text{Law}(\bar{X}_t^i) = \rho_t$ for each $i = 1, \dots, N$. Such system is coupled to (3.1) exclusively through the shared initial data $(X_0^i)_{i=1,\dots,N}$ as well as Brownian motion paths $((B_t^i)_{t \geq 0})_{i=1,\dots,N}$.

Pursuing a stability-like estimate between the interacting particle system $(X^i)_{i=1,\dots,N}$ and its non-interacting counterpart $(\bar{X}^i)_{i=1,\dots,N}$, see, e.g., McKean's or Sznitman's proof of McKean's theorem [CD22b, Theorem 3.1], we seek to derive either a pointwise estimate of the form

$$\max_{i=1,\dots,N} \sup_{t \in [0, T]} \mathbb{E} \|X_t^i - \bar{X}_t^i\|_2^2 \leq CN^{-1}, \quad (3.53)$$

or the stronger pathwise result

$$\max_{i=1,\dots,N} \mathbb{E} \sup_{t \in [0, T]} \|X_t^i - \bar{X}_t^i\|_2^2 \leq CN^{-1}, \quad (3.54)$$

where in either case the constant C is independent of N . Despite the simplicity of the synchronous coupling, it typically yields in (3.53) or (3.54) a favorable convergence rate w.r.t. the number of particles N for any finite time horizon $T > 0$, see [CD22b, Theorems 3.1 and 3.3], however, typically suffers, due to a standard Grönwall argument, from an exponential dependence of the constant C on the time horizon T .

Due to a lack of global Lipschitz continuity of the drift and diffusion terms in (3.1) and (3.2), respectively, which impedes the application of McKean's theorem [CD22b, Theorem 3.1], the mean-field approximation of CBO as in (3.53) or (3.54) has been left as a difficult and open problem in [Car+18, Remark 3.3]. However, since then, several works, which we outline in what follows, have shed light on this issue, see also [CBO-I, Remark 2] and [GHV23, Section 1.3]. While the works [For+20; CBO-I; KST23; CBO-III; GHV23] described in (ii)–(vi) below rely on the afore-described synchronous coupling method and target a quantitative propagation of chaos result, the work [HQ22] summarized in (i) derives a qualitative statement through the stochastic tightness and compactness method.

- (i) The authors of [HQ22] employ a tightness and compactness argument in the path space. In a first step, they show that the sequence $\{\widehat{\rho}^N\}_{N \geq 2}$ of $\mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ -valued random variables, denoting the empirical random particle measure associated with the microscopic CBO dynamics (3.1), is tight. This permits to

employ Prokhorov's theorem to obtain, up to a subsequence, some limiting random measure. Consecutively, to identify this limit, they verify that it weakly satisfies the macroscopic Fokker-Planck equation (3.3), which is deterministic, hence showing that the limiting measure is actually deterministic. With this they show that, as $N \rightarrow \infty$, $\widehat{\rho}^N$ converges in law to the deterministic particle distribution $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, which satisfies the mean-field PDE (3.3). As mentioned before, this convergence is unfortunately only qualitative and does not allow to obtain an informative quantitative convergence rate with respect to the number of particles N . However, it closed the mean-field limit gap qualitatively.

- (ii) A desired quantitative result has been established for the first time in [For+20, Theorem 3.1 and Remark 3.1] for the CBO variant of the microscopic system (2.11), which is constrained to compact hypersurfaces Γ . By pursuing the coupling method using a synchronous coupling and exploiting the inherent compactness of the dynamics due to its confinement to Γ , which implies the global Lipschitz continuity of the consensus point, the authors derive the weak convergence of the continuous-time analog of the variant (2.11) to the corresponding mean-field limit in the sense that for all $\phi \in \mathcal{C}_b^1(\mathbb{R}^d)$ it holds

$$\sup_{t \in [0, T]} \mathbb{E} \left[\left| \langle \widehat{\rho}_t^N, \phi \rangle - \langle \rho_t, \phi \rangle \right|^2 \right] \leq \frac{C}{N} \|\phi\|_{\mathcal{C}^1(\mathbb{R}^d)}^2. \quad (3.55)$$

The obtained convergence rate reads $CN^{-1/2}$ with C depending in particular on

$$C_\alpha := \exp \left(\alpha \left(\sup_{x \in \Gamma} \mathcal{E}(x) - \inf_{x \in \Gamma} \mathcal{E}(x) \right) T \right). \quad (3.56)$$

- (iii) However, this left open the question about a quantitative mean-field approximation result for the dynamics (3.1) on the plane, to which we provided a first answer in [CBO-I, Section 3.3] by proving that pointwise propagation of chaos holds with the favorable convergence rate $N^{-1/2}$ in the number of particles N on a set of high probability. Using the coupling method via a synchronous coupling and leveraging the techniques from (ii) as well as the boundedness of moments established in [Car+18, Lemma 3.4], we establish in Proposition 3.16 a result about a quantitative mean-field approximation of the form (3.53) on a restricted set of bounded processes. For this set, on which the drift and diffusion terms are globally Lipschitz, we derive in Lemma 3.15 an estimate of its probability. Combining these two statements yields propagation of chaos with high probability. The details and technical steps are outlined in the remainder of this section.
- (iv) In the work [KST23], the coupling method is combined with the use of stopping times, introduced to handle the lack of global Lipschitz continuity of the dynamics. This allows to derive a quantitative pointwise mean-field approximation result. While the authors' statement [KST23, Theorem 4.2] is non-probabilistic, the obtained convergence rate scales as $\log(\log(N))^{-1/2}$ in the number of particles, which is suboptimal.

- (v) For the CBO variant (3.108) with truncated noise, which we investigate in [CBO-III] and describe concisely in Section 3.2, we obtain in Proposition 3.24 a non-probabilistic pointwise mean-field approximation result of the form (3.53) with favorable rate $N^{-1/2}$ by using a synchronous coupling. This result relies on the fact that the truncation in the noise term allows for a sufficient amount of boundedness, which in turn yields sub-Gaussianity of the coupled system, see Lemma 3.22.
- (vi) A more refined and detailed analysis compared to the ones mentioned in (iii) and (iv) is presented in the work [GHV23], which extends [Ger23]. The authors of [GHV23] derive a pathwise mean-field approximation result of the form (3.54) using the coupling method with a synchronous coupling, see [GHV23, Theorem 2.6]. Sznitman's classical argument for the proof of McKean's theorem is adapted with the intention of allowing coefficients that are not globally Lipschitz continuous. The key novelties include an improved stability estimate for the consensus point [Car+18, Lemmas 3.1 and 3.2], see [GHV23, Corollary 3.3] in comparison to Lemma 3.17, as well as relying on results from the statistics literature [DL09, Theorem 1] to obtain a sampling estimate, see [GHV23, Lemma 3.7] in comparison to Lemma 3.18. The only slightly stronger assumption required w.r.t. prior work is the higher moment bound $\rho_0 \in \mathcal{P}_6(\mathbb{R}^d)$ on the initial measure. [GHV23] additionally proves the mean-field approximation for CBS [Car+22].

Although the work [GHV23] summarized in (vi) closes the question about a mean-field approximation for the CBO dynamics (3.1), it should be emphasized once more that the constant in (3.54) depends in all described cases exponentially on the time horizon T . To remedy this and to obtain uniform-in-time estimates [Dur+20], more involved coupling strategies and different metrics are necessary.

In what follows, however, let us present in more detail the mean-field approximation result (iii), which we put forward in [CBO-I, Section 3.3], however, by adapting it to the setting of anisotropic noise. For this purpose, let us introduce the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ over which all considered stochastic processes get their realizations, and define a subset Ω_M of Ω of suitably bounded processes according to

$$\Omega_M := \left\{ \omega \in \Omega : \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max \left\{ \|X_t^i(\omega)\|_2^4, \|\bar{X}_t^i(\omega)\|_2^4 \right\} \leq M \right\}. \quad (3.57)$$

Throughout this section, $M > 0$ denotes a constant which we shall adjust at the end of the proof of Theorem 3.19. Before stating the mean-field approximation result, Proposition 3.16, let us estimate the measure of the set Ω_M in Lemma 3.15.

Lemma 3.15 (Moment bounds for CBO, cf. [CBO-I, Lemma 15]). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy W1–W2. Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N \in \mathbb{N}$ be fixed. Moreover, let $((X_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote the strong solution to system (3.1) and let $((\bar{X}_t^i)_{t \geq 0})_{i=1, \dots, N}$ be N independent copies of the strong solution to the mean-field dynamics (3.2). Then,

for any $M > 0$ we have

$$\mathbb{P}(\Omega_M) = \mathbb{P}\left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max\{\|X_t^i\|_2^4, \|\bar{X}_t^i\|_2^4\} \leq M\right) \geq 1 - \frac{2K}{M}, \quad (3.58)$$

where $K = K(\lambda, \sigma, d, T, b_1, b_2)$ is a constant, which is in particular independent of N . For the constants b_1 and b_2 we have

$$b_1 = 0 \quad \text{and} \quad b_2 = e^{\alpha(\bar{\varepsilon} - \varepsilon)} \quad (3.59)$$

in case the first condition of [W2](#) holds and

$$b_1 = C_4^2 + b_2 \quad \text{and} \quad b_2 = 2 \frac{C_2}{C_3} \left(1 + \frac{1}{\alpha C_3} \frac{1}{C_4^2}\right) \quad (3.60)$$

in case of the second condition of [W2](#), see also [[Car+18](#), Lemma 3.3].

Proof. By combining the ideas of [[Car+18](#), Lemma 3.4] with a Doob-like inequality, we derive a bound for $\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max\{\|X_t^i\|_2^4, \|\bar{X}_t^i\|_2^4\}$, which ensures that $\hat{\rho}_t^N, \bar{\rho}_t^N \in \mathcal{P}_4(\mathbb{R}^d)$ with high probability. Here, $\bar{\rho}^N$ denotes the empirical measure associated with the processes $(\bar{X}^i)_{i=1, \dots, N}$.

By employing the inequality $(z + z')^q \leq 2^{q-1}(z^q + z'^q)$, $q \geq 1$ we note that

$$\begin{aligned} \|X_t^i\|_2^{2p} &\leq 2^{2p-1} \|X_0^i\|_2^{2p} + 2^{2(2p-1)} \lambda^{2p} \left\| \int_0^t (X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)) d\tau \right\|_2^{2p} \\ &\quad + 2^{2(2p-1)} \sigma^{2p} \left\| \int_0^t D(X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)) dB_\tau^i \right\|_2^{2p} \end{aligned} \quad (3.61)$$

for all $i = 1, \dots, N$. Taking first the supremum over $t \in [0, T]$ and consecutively the expectation on both sides of the former inequality yields

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \|X_t^i\|_2^{2p} &\leq 2^{2p-1} \mathbb{E} \|X_0^i\|_2^{2p} + 2^{2(2p-1)} \lambda^{2p} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t (X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)) d\tau \right\|_2^{2p} \\ &\quad + 2^{2(2p-1)} \sigma^{2p} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t D(X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)) dB_\tau^i \right\|_2^{2p}. \end{aligned} \quad (3.62)$$

The second term on the right-hand side of (3.62) can be further bounded by

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t (X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)) d\tau \right\|_2^{2p} &\leq \max\{1, T^{2p-1}\} \mathbb{E} \sup_{t \in [0, T]} \int_0^t \|X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)\|_2^{2p} d\tau \\ &\leq \max\{1, T^{2p-1}\} \mathbb{E} \int_0^T \|X_\tau^i - x_\alpha^\varepsilon(\hat{\rho}_\tau^N)\|_2^{2p} d\tau \end{aligned} \quad (3.63)$$

as a consequence of Jensen's inequality. For the third term on the right-hand side of (3.62) we first note that the expression $\int_0^t D(X_\tau^i - x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)) dB_\tau^i$ is a martingale. This is due to [Øks03, Corollary 3.2.6] since its expected quadratic variation is finite as required by [Øks03, Definition 3.1.4]. The latter immediately follows from the regularity established in [Car+18, Lemma 3.4]. Therefore we can apply the Burkholder-Davis-Gundy inequality [RY99, Chapter IV, Theorem 4.1], which gives for a generic constant C_{2p} the bound

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t D(X_\tau^i - x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)) dB_\tau^i \right\|_2^{2p} &\leq C_{2p} \sup_{t \in [0, T]} \mathbb{E} \left(\int_0^t \|X_\tau^i - x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)\|_2^2 d\tau \right)^p \\ &\leq C_{2p} \max\{1, T^{p-1}\} \mathbb{E} \int_0^T \|X_\tau^i - x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)\|_2^{2p} d\tau. \end{aligned} \quad (3.64)$$

Here, the latter step is again due to Jensen's inequality. The right-hand sides of (3.63) and (3.64) can be further bounded since

$$\begin{aligned} \mathbb{E} \int_0^T \|X_\tau^i - x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)\|_2^{2p} d\tau &\leq 2^{2p-1} \mathbb{E} \int_0^T \left(\|X_\tau^i\|_2^{2p} + \|x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)\|_2^{2p} \right) d\tau \\ &\leq 2^{2p-1} \mathbb{E} \int_0^T \left(\|X_\tau^i\|_2^{2p} + 2^{p-1} \left(b_1^p + b_2^p \int \|x\|_2^{2p} d\widehat{\rho}_\tau^N(x) \right) \right) d\tau, \end{aligned} \quad (3.65)$$

where in the last step we made use of [Car+18, Lemma 3.3], which shows that

$$\|x_\alpha^\mathcal{E}(\widehat{\rho}_\tau^N)\|_2^2 \leq \int \|x\|_2^2 \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\widehat{\rho}_\tau^N)}} d\widehat{\rho}_\tau^N(x) \leq b_1 + b_2 \int \|x\|_2^2 d\widehat{\rho}_\tau^N(x), \quad (3.66)$$

with $b_1 = 0$ and $b_2 = e^{\alpha(\bar{\mathcal{E}} - \underline{\mathcal{E}})}$ in the case that \mathcal{E} is bounded, and

$$b_1 = C_4^2 + b_2 \quad \text{and} \quad b_2 = 2 \frac{C_2}{C_3} \left(1 + \frac{1}{\alpha C_3} \frac{1}{C_4^2} \right) \quad (3.67)$$

in the case that \mathcal{E} satisfies the coercivity assumption (3.9). Inserting the upper bounds (3.63) and (3.64) together with the estimate (3.65) into (3.62) yields

$$\mathbb{E} \sup_{t \in [0, T]} \|X_t^i\|_2^{2p} \leq C \left(1 + \mathbb{E} \|X_0^i\|_2^{2p} + \mathbb{E} \int_0^T \|X_\tau^i\|_2^{2p} + \int \|x\|_2^{2p} d\widehat{\rho}_\tau^N(x) d\tau \right) \quad (3.68)$$

with a constant $C = C(p, \lambda, \sigma, T, b_1, b_2)$. Averaging (3.68) over i allows to bound

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \int \|x\|_2^{2p} d\widehat{\rho}_t^N(x) &\leq C \left(1 + \mathbb{E} \int \|x\|_2^{2p} d\widehat{\rho}_0^N(x) + 2 \int_0^T \mathbb{E} \int \|x\|_2^{2p} d\widehat{\rho}_\tau^N(x) d\tau \right) \\ &\leq C \left(1 + \mathbb{E} \int \|x\|_2^{2p} d\widehat{\rho}_0^N(x) + 2 \int_0^T \mathbb{E} \sup_{\tau' \in [0, \tau]} \int \|x\|_2^{2p} d\widehat{\rho}_{\tau'}^N(x) d\tau \right), \end{aligned} \quad (3.69)$$

which ensures that $\mathbb{E} \sup_{t \in [0, T]} \int \|x\|_2^{2p} d\widehat{\rho}_t^N(x)$ is bounded independently of N by Grönwall's inequality provided $\rho_0 \in \mathcal{P}_{2p}(\mathbb{R}^d)$. Since this holds by the assumption $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ for $p = 2$, there exists a constant $K = K(\lambda, \sigma, T, b_1, b_2)$, in particular independently of N , such that $\mathbb{E} \sup_{t \in [0, T]} \int \|x\|_2^4 d\widehat{\rho}_t^N(x) \leq K$.

Following analogous arguments for \bar{X}_t^i allows to derive

$$\mathbb{E} \sup_{t \in [0, T]} \|\bar{X}_t^i\|_2^{2p} \leq C \left(1 + \mathbb{E} \|\bar{X}_0^i\|_2^{2p} + \mathbb{E} \int_0^T \|\bar{X}_\tau^i\|_2^{2p} + \int \|x\|_2^{2p} d\rho_\tau(x) d\tau \right) \quad (3.70)$$

in place of (3.68). Noticing that $\int \|x\|_2^{2p} d\rho_\tau(x) = \mathbb{E} \|\bar{X}_\tau^i\|_2^{2p}$ for all $\tau \in [0, T]$ and averaging the latter over i directly permits to prove $\mathbb{E} \sup_{t \in [0, T]} \int \|x\|_2^{2p} d\bar{\rho}_t(x) \leq K$ by applying Grönwall's inequality, again provided that $\rho_0 \in \mathcal{P}_{2p}(\mathbb{R}^d)$. With this being the case for $p = 2$ and by choosing K sufficiently large for either estimate, the statement follows from a union bound and Markov's inequality. More precisely,

$$\begin{aligned} & \mathbb{P} \left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max \left\{ \|X_t^i\|_2^4, \|\bar{X}_t^i\|_2^4 \right\} > M \right) \\ & \leq \mathbb{P} \left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|X_t^i\|_2^4 > M \right) + \mathbb{P} \left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|\bar{X}_t^i\|_2^4 > M \right) \\ & \leq \frac{\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|X_t^i\|_2^4}{M} + \frac{\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|\bar{X}_t^i\|_2^4}{M} \\ & \leq 2 \frac{K}{M}, \end{aligned} \quad (3.71)$$

which concludes the proof. \square

Lemma 3.15 proves that the processes are bounded with high probability uniformly in time. Therefore, by restricting the analysis to Ω_M , we can obtain the following quantitative mean-field approximation result by employing the coupling method [CD22a; CD22b] using a synchronous coupling between the stochastic processes X^i and \bar{X}^i , see, e.g., [CD22b, Section 4.1.2].

Proposition 3.16 (Probabilistic mean-field approximation of CBO, cf. [CBO-I, Proposition 16]). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy W1–W2. Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N \in \mathbb{N}$ be fixed. Moreover, let $((X_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote the strong solution to system (3.1) and let $((\bar{X}_t^i)_{t \geq 0})_{i=1, \dots, N}$ be N independent copies of the strong solution to the mean-field dynamics (3.2). If $(X_t^i)_{t \geq 0}$ and $(\bar{X}_t^i)_{t \geq 0}$ share the initial data as well as the Brownian motion paths $(B_t^i)_{t \geq 0}$ for all $i = 1, \dots, N$, then we have

$$\max_{i=1, \dots, N} \sup_{t \in [0, T]} \mathbb{E} \left[\left\| X_t^i - \bar{X}_t^i \right\|_2^2 \mid \Omega_M \right] \leq C_{\text{MFA}} N^{-1} \quad (3.72)$$

with $C_{\text{MFA}} = C_{\text{MFA}}(\alpha, \lambda, \sigma, T, C_1, C_2, M, K, \mathcal{M}_2, b_1, b_2)$, where K is as in Lemma 3.15 and \mathcal{M}_2 denotes a second-order moment bound of ρ .

Before providing the proof of [Proposition 3.16](#) we require two auxiliary results. For this we define the cutoff function I_M according to

$$I_M(t) = \begin{cases} 1, & \text{if } \frac{1}{N} \sum_{i=1}^N \max \left\{ \|X_\tau^i\|_2^4, \|\bar{X}_\tau^i\|_2^4 \right\} \leq M, \text{ for all } \tau \in [0, t], \\ 0, & \text{else,} \end{cases} \quad (3.73)$$

which is adapted to the natural filtration and has the property $I_M(t) = I_M(t)I_M(\tau)$ for all $\tau \in [0, t]$.

Lemma 3.17 (Stability estimate for the consensus point, [CBO-I, Lemma 25]). Let I_M be as defined in (3.73). Then, under [Assumption 3.3](#), it holds

$$\left\| x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N) \right\|_2^2 I_M(\tau) \leq C \frac{1}{N} \sum_{i=1}^N \left\| X_\tau^i - \bar{X}_\tau^i \right\|_2^2 I_M(\tau) \quad (3.74)$$

for a constant $C = C(\alpha, C_1, C_2, M)$.

Proof. The proof follows the steps taken in [[Car+18](#), Lemmas 3.1 and 3.2].

Let us first note that by exploiting that the quantity $\frac{1}{N} \sum_{i=1}^N \|X_\tau^i\|_2^4$ is bounded uniformly by M due to the multiplication with $I_M(\tau)$, we obtain with Jensen's inequality that

$$\begin{aligned} \frac{e^{-\alpha\mathcal{E}} I_M(\tau)}{\frac{1}{N} \sum_{i=1}^N \omega_\alpha^\mathcal{E}(X_\tau^i)} &\leq \frac{I_M(\tau)}{\exp\left(-\alpha \frac{1}{N} \sum_{i=1}^N (\mathcal{E}(X_\tau^i) - \mathcal{E})\right)} \\ &\leq \frac{I_M(\tau)}{\exp\left(-\alpha C_2 \frac{1}{N} \sum_{i=1}^N (1 + \|X_\tau^i\|_2^2)\right)} \\ &\leq \exp\left(\alpha C_2 (1 + \sqrt{M})\right) =: c_M, \end{aligned} \quad (3.75)$$

where, in the second inequality, we used the assumption (3.8) on \mathcal{E} . An analogous statement can be obtained for the processes \bar{X}_τ^i .

For the norm of the difference between $x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N)$ and $x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N)$ we have the decomposition

$$\begin{aligned} \left\| x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N) \right\|_2 I_M(\tau) &= \left\| \frac{\sum_{i=1}^N X_\tau^i \omega_\alpha^\mathcal{E}(X_\tau^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} - \frac{\sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} \right\|_2 I_M(\tau) \\ &\leq (\|T_1\|_2 + \|T_2\|_2 + \|T_3\|_2) I_M(\tau), \end{aligned} \quad (3.76)$$

where the terms T_1 , T_2 and T_3 are obtained by inserting mixed terms with respect to X_τ^i and \bar{X}_τ^i . They are defined implicitly below and their norm is bounded as follows. For

the first term T_1 we have

$$\begin{aligned}
 \|T_1\|_2 I_M(\tau) &= \left\| \frac{1}{N} \sum_{i=1}^N (X_\tau^i - \bar{X}_\tau^i) \frac{\omega_\alpha^\mathcal{E}(X_\tau^i)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} \right\|_2 I_M(\tau) \\
 &\leq \frac{1}{N} \sum_{i=1}^N \|X_\tau^i - \bar{X}_\tau^i\|_2 \left| \frac{\omega_\alpha^\mathcal{E}(X_\tau^i)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} \right| I_M(\tau) \\
 &\leq \left| \frac{e^{-\alpha\mathcal{E}} I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} \right| \frac{1}{N} \sum_{i=1}^N \|X_\tau^i - \bar{X}_\tau^i\|_2 I_M(\tau) \\
 &\leq c_M \sqrt{\frac{1}{N} \sum_{i=1}^N \|X_\tau^i - \bar{X}_\tau^i\|_2^2} I_M(\tau),
 \end{aligned} \tag{3.77}$$

where we made use of (3.75) and Cauchy-Schwarz inequality in the last step. For the second term T_2 , by using the assumption (3.7) on \mathcal{E} in the third line and by following similar steps, we obtain

$$\begin{aligned}
 \|T_2\|_2 I_M(\tau) &= \left\| \frac{1}{N} \sum_{i=1}^N (\omega_\alpha^\mathcal{E}(X_\tau^i) - \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)) \frac{\bar{X}_\tau^i}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} I_M(\tau) \right\|_2 I_M(\tau) \\
 &\leq \frac{1}{N} \sum_{i=1}^N |\omega_\alpha^\mathcal{E}(X_\tau^i) - \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)| \left\| \frac{\bar{X}_\tau^i}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} \right\|_2 I_M(\tau) \\
 &\leq \alpha C_1 e^{-\alpha\mathcal{E}} \frac{1}{N} \sum_{i=1}^N (\|X_\tau^i\|_2 + \|\bar{X}_\tau^i\|_2) \|X_\tau^i - \bar{X}_\tau^i\|_2 \\
 &\quad \cdot \frac{\|\bar{X}_\tau^i\|_2}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} I_M(\tau) \\
 &\leq \frac{3}{2} \alpha C_1 \left| \frac{e^{-\alpha\mathcal{E}} I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} \right| \frac{1}{N} \sum_{i=1}^N (\|X_\tau^i\|_2^2 + \|\bar{X}_\tau^i\|_2^2) \|X_\tau^i - \bar{X}_\tau^i\|_2 I_M(\tau) \\
 &\leq \frac{3}{2} \alpha C_1 c_M \sqrt{\frac{1}{N} \sum_{i=1}^N (\|X_\tau^i\|_2^4 + \|\bar{X}_\tau^i\|_2^4)} I_M(\tau) \\
 &\quad \cdot \sqrt{\frac{1}{N} \sum_{i=1}^N \|X_\tau^i - \bar{X}_\tau^i\|_2^2} I_M(\tau) \\
 &\leq 3\alpha C_1 c_M M^{\frac{1}{2}} \sqrt{\frac{1}{N} \sum_{i=1}^N \|X_\tau^i - \bar{X}_\tau^i\|_2^2} I_M(\tau).
 \end{aligned} \tag{3.78}$$

Analogously, for the third term T_3 , we get

$$\begin{aligned}
 \|T_3\|_2 I_M(\tau) &= \left\| \frac{\sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)} - \frac{\sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} \right\|_2 I_M(\tau) \\
 &\leq \frac{1}{N} \sum_{j=1}^N \left| \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j) - \omega_\alpha^\mathcal{E}(X_\tau^j) \right| \left\| \frac{\frac{1}{N} \sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)}{\left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)\right) \left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)\right)} \right\|_2 I_M(\tau) \\
 &\leq \alpha C_1 e^{-2\alpha\mathcal{E}} \frac{1}{N} \sum_{j=1}^N \left(\|X_\tau^j\|_2 + \|\bar{X}_\tau^j\|_2 \right) \|X_\tau^j - \bar{X}_\tau^j\|_2 I_M(\tau) \\
 &\quad \cdot \frac{\frac{1}{N} \sum_{i=1}^N \|\bar{X}_\tau^i\|_2 I_M(\tau)}{\left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(X_\tau^j)\right) \left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)\right)} \\
 &\leq \alpha C_1 c_M^2 M^{\frac{1}{4}} \frac{1}{N} \sum_{i=1}^N \left(\|X_\tau^i\|_2 + \|\bar{X}_\tau^i\|_2 \right) \|X_\tau^i - \bar{X}_\tau^i\|_2 I_M(\tau) \\
 &\leq \sqrt{2} \alpha C_1 c_M^2 M^{\frac{1}{4}} \sqrt{\frac{1}{N} \sum_{j=1}^N \left(\|X_\tau^j\|_2^2 + \|\bar{X}_\tau^j\|_2^2 \right)} I_M(\tau) \\
 &\quad \cdot \sqrt{\frac{1}{N} \sum_{j=1}^N \|X_\tau^j - \bar{X}_\tau^j\|_2^2 I_M(\tau)} \\
 &\leq 2\alpha C_1 c_M^2 M^{\frac{1}{2}} \sqrt{\frac{1}{N} \sum_{j=1}^N \|X_\tau^j - \bar{X}_\tau^j\|_2^2 I_M(\tau)}.
 \end{aligned} \tag{3.79}$$

By inserting the three individual bounds (3.77), (3.78) and (3.79) into (3.76) and taking the squares of both sides, we obtain the upper bound from the statement. \square

Lemma 3.18 (Sampling estimate for the consensus point, [CBO-I, Lemma 26]). Let $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and let I_M be as defined in (3.73). Then, under Assumption 3.3, it holds

$$\sup_{\tau \in [0, T]} \mathbb{E} \left\| x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau) \right\|_2^2 I_M(\tau) \leq CN^{-1} \tag{3.80}$$

for a constant $C = C(\alpha, C_2, M, \mathcal{M}_2, b_1, b_2)$, where \mathcal{M}_2 denotes the second-order moment bound of ρ and where b_1 and b_2 are the problem-dependent constants specified in (3.67).

Proof. The proof follows the steps taken in [For+20, Lemma 3.1].

By inserting a mixed term, we can bound the norm of the difference between $x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N)$ and $x_\alpha^\mathcal{E}(\rho_\tau)$ by

$$\begin{aligned} \left\| x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau) \right\|_2 I_M(\tau) &= \left\| \sum_{i=1}^N \bar{X}_\tau^i \frac{\omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} - \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_\tau)}} d\rho_\tau(x) \right\|_2 I_M(\tau) \\ &\leq (\|T_1\|_2 + \|T_2\|_2) I_M(\tau), \end{aligned} \quad (3.81)$$

where the terms T_1 and T_2 are defined implicitly and bounded in what follows. By utilizing the bound (3.75), for the first term T_1 , we get

$$\begin{aligned} \|T_1\|_2 I_M(\tau) &= \left\| \sum_{i=1}^N \bar{X}_\tau^i \frac{\omega_\alpha^\mathcal{E}(\bar{X}_\tau^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} - \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} d\rho_\tau(x) \right\|_2 I_M(\tau) \\ &= \left| \frac{I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} \right| \left\| \frac{1}{N} \sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int x \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right\|_2 \\ &\leq c_M e^{\alpha\mathcal{E}} \left\| \frac{1}{N} \sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int x \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right\|_2. \end{aligned} \quad (3.82)$$

Similarly, for the second term we have

$$\begin{aligned} \|T_2\|_2 I_M(\tau) &= \left\| \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} d\rho_\tau(x) - \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_\tau)}} d\rho_\tau(x) \right\|_2 I_M(\tau) \\ &= \left| \frac{I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j)} \right| \left\| x_\alpha^\mathcal{E}(\rho_\tau) \right\|_2 \left| \frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j) - \|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_\tau)} \right| \\ &\leq c_M e^{\alpha\mathcal{E}} \sqrt{b_1 + b_2 \mathcal{M}_2} \left| \frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^j) - \int \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right|, \end{aligned} \quad (3.83)$$

where the last step uses that by Jensen's inequality and [Car+18, Lemma 3.3] it holds

$$\left\| x_\alpha^\mathcal{E}(\rho_\tau) \right\|_2^2 \leq \int \|x\|_2^2 \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\rho_\tau)}} d\rho_\tau(x) \leq b_1 + b_2 \int \|x\|_2^2 d\rho_\tau(x) \leq b_1 + b_2 \mathcal{M}_2 \quad (3.84)$$

with constants b_1 and b_2 as specified in (3.67) and \mathcal{M}_2 denoting a bound on the second-order moment of ρ , which exists according to the regularity of ρ established in [CBO-II, Theorem 1] as a consequence of the initial regularity $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$. In order to further bound (3.82) and (3.83), respectively, let us introduce the random variables

$$Z_\tau^i = \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int x \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \quad (3.85)$$

and

$$z_\tau^i = \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x), \quad (3.86)$$

which have zero expectation, i.e., $\mathbb{E}Z_\tau^i = 0$ and $\mathbb{E}z_\tau^i = 0$. Moreover, we observe that

$$\frac{1}{N} \sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int x \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) = \frac{1}{N} \sum_{i=1}^N Z_\tau^i \quad (3.87)$$

and

$$\frac{1}{N} \sum_{i=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) = \frac{1}{N} \sum_{i=1}^N z_\tau^i, \quad (3.88)$$

respectively. Moreover, due to the independence of the \bar{X}_τ^i 's the Z_τ^i 's are independent and thus satisfy $\mathbb{E}Z_\tau^i Z_\tau^j = 0$ for $i \neq j$. Using this we can rewrite

$$\begin{aligned} \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \bar{X}_\tau^i \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int x \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right\|_2^2 &= \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N Z_\tau^i \right\|_2^2 = \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \sum_{j=1}^N \langle Z_\tau^i, Z_\tau^j \rangle \\ &= \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \|Z_\tau^i\|_2^2 = \frac{1}{N} \mathbb{E} \|Z_\tau^1\|_2^2 \\ &\leq 4e^{-2\alpha\mathcal{E}} \mathcal{M}_2 \frac{1}{N}, \end{aligned} \quad (3.89)$$

where the inequality in the last step is due to the estimate

$$\begin{aligned} \mathbb{E} \|Z_\tau^1\|_2^2 &\leq 2\mathbb{E} \|\bar{X}_\tau^1 \omega_\alpha^\mathcal{E}(\bar{X}_\tau^1)\|_2^2 + 2 \left\| \int x \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right\|_2^2 \\ &\leq 2e^{-2\alpha\mathcal{E}} \left(\mathbb{E} \|\bar{X}_\tau^1\|_2^2 + \int \|x\|_2^2 d\rho_\tau(x) \right) \\ &\leq 4e^{-2\alpha\mathcal{E}} \mathcal{M}_2. \end{aligned} \quad (3.90)$$

Following analogous arguments and noting that

$$\begin{aligned} \mathbb{E} |z_\tau^1|^2 &\leq 2\mathbb{E} |\omega_\alpha^\mathcal{E}(\bar{X}_\tau^1)|^2 + 2 \left| \int \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right|^2 \\ &\leq 4e^{-2\alpha\mathcal{E}} \end{aligned} \quad (3.91)$$

yields the inequality

$$\begin{aligned} \mathbb{E} \left| \frac{1}{N} \sum_{i=1}^N \omega_\alpha^\mathcal{E}(\bar{X}_\tau^i) - \int \omega_\alpha^\mathcal{E}(x) d\rho_\tau(x) \right|^2 &= \frac{1}{N} \mathbb{E} |z_\tau^1|^2 \\ &\leq 4e^{-2\alpha\mathcal{E}} \frac{1}{N}. \end{aligned} \quad (3.92)$$

The statement now follows by combining the two individual bounds (3.89) and (3.92) with (3.81) after taking the square and expectation there. \square

We now have the necessary technical tools at hand to provide the proof of [Proposition 3.16](#).

Proof of Proposition 3.16. By exploiting the boundedness of the dynamics established in [Lemma 3.15](#) through a cutoff technique, we can follow the steps taken in [[For+20](#), Theorem 3.1].

Let us again define the cutoff function I_M as in [\(3.73\)](#). This allows us to obtain for $\mathbb{E}\|X_t^i - \bar{X}_t^i\|_2^2 I_M(t)$ the inequality

$$\begin{aligned}
 \mathbb{E}\|X_t^i - \bar{X}_t^i\|_2^2 I_M(t) &\leq 2\mathbb{E}\|X_0^i - \bar{X}_0^i\|_2^2 \\
 &\quad + 4\lambda^2\mathbb{E}\left\|\int_0^t \left((X_\tau^i - x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N)) - (\bar{X}_\tau^i - x_\alpha^\mathcal{E}(\rho_\tau))\right) I_M(\tau) d\tau\right\|_2^2 \\
 &\quad + 4\sigma^2\mathbb{E}\left\|\int_0^t \left(D(X_\tau^i - x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N)) - D(\bar{X}_\tau^i - x_\alpha^\mathcal{E}(\rho_\tau))\right) I_M(\tau) dB_\tau^i\right\|_2^2 \\
 &\leq 2\mathbb{E}\|X_0^i - \bar{X}_0^i\|_2^2 \\
 &\quad + 8\lambda^2 T \mathbb{E}\int_0^t \left(\|X_\tau^i - \bar{X}_\tau^i\|_2^2 + \|x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau)\|_2^2\right) I_M(\tau) d\tau \\
 &\quad + 8\sigma^2 \mathbb{E}\int_0^t \left(\|X_\tau^i - \bar{X}_\tau^i\|_2^2 + \|x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau)\|_2^2\right) I_M(\tau) d\tau,
 \end{aligned} \tag{3.93}$$

where we used in the first step that the processes X_τ^i and \bar{X}_τ^i share the Brownian motion paths, and in the second step both Itô isometry and Jensen's inequality. Noting further that the processes also share the initial data, we are left with

$$\mathbb{E}\|X_t^i - \bar{X}_t^i\|_2^2 I_M(t) \leq 8(\lambda^2 T + \sigma^2) \int_0^t \mathbb{E}\left(\|X_\tau^i - \bar{X}_\tau^i\|_2^2 + \|x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau)\|_2^2\right) I_M(\tau) d\tau, \tag{3.94}$$

where it remains to control $\mathbb{E}\|x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau)\|_2^2 I_M(\tau)$. By means of [Lemmas 3.17](#) and [3.18](#) below we have the bound

$$\begin{aligned}
 \mathbb{E}\|x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau)\|_2^2 I_M(\tau) &\leq 2\mathbb{E}\|x_\alpha^\mathcal{E}(\hat{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N)\|_2^2 I_M(\tau) \\
 &\quad + 2\mathbb{E}\|x_\alpha^\mathcal{E}(\bar{\rho}_\tau^N) - x_\alpha^\mathcal{E}(\rho_\tau)\|_2^2 I_M(\tau) \\
 &\leq C \left(\frac{1}{N} \sum_{i=1}^N \mathbb{E}\|X_\tau^i - \bar{X}_\tau^i\|_2^2 I_M(\tau) + N^{-1} \right) \\
 &\leq C \left(\max_{i=1,\dots,N} \mathbb{E}\|X_\tau^i - \bar{X}_\tau^i\|_2^2 I_M(\tau) + N^{-1} \right)
 \end{aligned} \tag{3.95}$$

for a constant $C = C(\alpha, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. After integrating the bound [\(3.95\)](#) into [\(3.94\)](#) and taking the maximum over i we are left with

$$\max_{i=1,\dots,N} \mathbb{E}\|X_t^i - \bar{X}_t^i\|_2^2 I_M(t) \leq C \int_0^t \max_{i=1,\dots,N} \mathbb{E}\|X_\tau^i - \bar{X}_\tau^i\|_2^2 I_M(\tau) d\tau + CTN^{-1}, \tag{3.96}$$

where C depends additionally on λ , σ and T , i.e., $C = C(\alpha, \lambda, \sigma, T, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. The second part of the statement now follows from an application of Grönwall's inequality and by noting that $\mathbb{1}_{\Omega_M} \leq I_M(t)$ pointwise and for all $t \in [0, T]$. \square

3.1.3. Global Convergence in Probability: On Holistic Global Convergence Guarantees¹⁶

A combination of the results of the former two sections about the convergence in mean-field law, [Theorem 3.6](#) from [Section 3.1.1](#) and the quantitative mean-field approximation, [Proposition 3.16](#) from [Section 3.1.2](#), together with classical results of numerical approximation of SDEs [[Pla99](#)], allows us to obtain a probabilistic statement about the global convergence of CBO.

Theorem 3.19 (CBO converges globally, cf. [[CBO-I, Theorem 13](#)]). Fix $\varepsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Then, under the assumptions of [Theorem 3.6](#) and [Proposition 3.16](#), and with $K := T/\Delta t$, where T is as in [\(3.29\)](#) and for suitable $\Delta t > 0$, the iterations $((X_k^i)_{k=1, \dots, K})_{i=1, \dots, N}$ generated by the numerical scheme [\(2.2\)](#) converge in probability to x^* . More precisely, the empirical mean of the final iterations fulfills the quantitative error estimate

$$\left\| \frac{1}{N} \sum_{i=1}^N X_K^i - x^* \right\|_2^2 \leq \varepsilon_{\text{total}} \quad (3.97)$$

with probability larger than

$$1 - \left(\delta + \varepsilon_{\text{total}}^{-1} (6C_{\text{NA}}(\Delta t)^{2m} + 3C_{\text{MFA}}N^{-1} + 12\varepsilon) \right). \quad (3.98)$$

Here, m denotes the order of accuracy of the numerical scheme (for the Euler-Maruyama scheme $m = 1/2$) and ε is the error from [Theorem 3.6](#). Moreover, besides problem-dependent constants, $C_{\text{NA}} > 0$ depends linearly on the dimension d and the number of particles N , exponentially on the time horizon T , and on δ^{-1} ; $C_{\text{MFA}} > 0$ depends exponentially on the parameters α , λ and σ , on T , and on δ^{-1} .

Proof. We have the error decomposition¹⁷

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N X_K^i - x^* \right\|_2^2 \middle| \Omega_M \right] &\leq 3\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (X_K^i - X_{T^*}^i) \right\|_2^2 \middle| \Omega_M \right] \\ &+ 3\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (X_{T^*}^i - \bar{X}_{T^*}^i) \right\|_2^2 \middle| \Omega_M \right] + \frac{3}{1-\delta} \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \bar{X}_{T^*}^i - x^* \right\|_2^2, \end{aligned} \quad (3.99)$$

¹⁶In this section, we follow [[CBO-I, Section 3.3](#)] adapted to the setting of anisotropic noise [[CBO-II](#)].

¹⁷Let us remind the reader of the slight abuse of notations mentioned in footnote⁸ by using the same notation X^i for solutions to [\(2.2\)](#) as well as [\(3.1\)](#).

which divides the overall error into an approximation error of the numerical scheme, the mean-field approximation error and the optimization error in the mean-field limit. The first term on the right-hand side of (3.99) can be estimated by applying classical results about the convergence of numerical schemes for SDEs [Pla99] yielding

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (X_K^i - X_{T^*}^i) \right\|_2^2 \middle| \Omega_M \right] \leq C_{\text{NA}} (\Delta t)^{2m}. \quad (3.100)$$

The second term can be bounded by using precisely the quantitative mean-field approximation in form of Proposition 3.16, which establishes

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (X_{T^*}^i - \bar{X}_{T^*}^i) \right\|_2^2 \middle| \Omega_M \right] \leq C_{\text{MFA}} N^{-1}. \quad (3.101)$$

For the third term Theorem 3.6 gives after an application of Jensen's inequality

$$\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \bar{X}_{T^*}^i - x^* \right\|_2^2 \leq 2\mathcal{V}(\rho_{T^*}) \leq 2\varepsilon. \quad (3.102)$$

Combining these individual bounds with (3.99) allows to obtain the error estimate

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N X_K^i - x^* \right\|_2^2 \middle| \Omega_M \right] \leq 6C_{\text{NA}} (\Delta t)^{2m} + 3C_{\text{MFA}} N^{-1} + 12\varepsilon. \quad (3.103)$$

Let us now denote by $K_{\varepsilon_{\text{total}}}^N \subset \Omega$ the set, where (3.97) does not hold. Then we can estimate

$$\begin{aligned} \mathbb{P} \left(K_{\varepsilon_{\text{total}}}^N \right) &= \mathbb{P} \left(K_{\varepsilon_{\text{total}}}^N \cap \Omega_M \right) + \mathbb{P} \left(K_{\varepsilon_{\text{total}}}^N \cap \Omega_M^c \right) \\ &\leq \mathbb{P} \left(K_{\varepsilon_{\text{total}}}^N \middle| \Omega_M \right) \mathbb{P}(\Omega_M) + \mathbb{P}(\Omega_M^c) \\ &\leq \mathbb{P} \left(K_{\varepsilon_{\text{total}}}^N \middle| \Omega_M \right) + \frac{2K}{M} \\ &\leq \varepsilon_{\text{total}}^{-1} \left(6C_{\text{NA}} (\Delta t)^{2m} + 3C_{\text{MFA}} N^{-1} + 12\varepsilon \right) + \delta, \end{aligned}$$

where in the last step we employ the conditional version of Markov's inequality together with (3.103) to bound the first term. For the second it suffices to choose the M from (3.58) large enough. \square

Let us conclude this section about a review of our works [CBO-I; CBO-II] with some pivotal remarks related to the hardness of nonconvex optimization problems and how they are encoded in CBO.

Remark 3.20 (Weight/temperature parameter α , cf. [CBO-I, Section 3.3]). As a consequence of Theorem 3.6, the hardness of any optimization problem is necessar-

ily encoded in the mean-field approximation, or, more precisely, in the way how the empirical measure $\hat{\rho}^N$ of the finite particle dynamics (3.1) is used to approximate the mean-field limit (3.3). Proposition 3.16 addresses precisely this question, ensuring that, with arbitrarily high probability, the finite particle dynamics (3.1) keeps close to the mean-field dynamics (3.2). Since the rate of this convergence is of favorable order $N^{-1/2}$ in the number of particles N , the hardness of the problem is fully captured by the constant C_{MFA} in (3.72), which does not depend explicitly on the dimension d . Therefore, the mean-field approximation is, in general, not affected by the curse of dimensionality. Nevertheless, as our assumptions on the objective function \mathcal{E} do not exclude the class of NP-hard problems, it cannot be expected that CBO solves any problem, howsoever hard, with polynomial complexity.

This is reflected by the exponential dependence of C_{MFA} on the parameter α and its possibly worst-case linear dependence on the dimension d , as we discuss in what follows. However, several numerical experiments [Car+21; For+21; For+22; CBO-II; CBO-IV; BGP23; Car+23] in high dimensions confirm that in typical applications CBO and its variants perform comparably to state-of-the-art methods without the necessity of an exponentially large amount of particles. As mentioned before, characterizing α_0 in more detail is crucial in view of the mean-field approximation result, Proposition 3.16. We did not precisely specify α_0 in Theorem 3.6 since it seems challenging to provide informative bounds in all generality. Following [CBO-I, Remark 24], however, we can devise an informal derivation for objectives \mathcal{E} that are locally L -Lipschitz continuous on a neighborhood $B_R^\infty(x^*)$ of the global minimizer x^* and satisfy the coercivity condition (3.26) of A2 globally for $\nu = 1/2$. For a parameter-dependent constant $c = c(\vartheta, \lambda, \sigma)$, we obtain

$$\alpha > \alpha_0 = \frac{-8d}{c^2\eta^2\varepsilon} \log \left(\frac{c}{2^{d+1}\sqrt{2d}} \rho_0 \left(B_{\min\{R, c^2\eta^2\varepsilon/(8dL)\}}^\infty(x^*) \right) \right) \quad (3.104)$$

provided that the probability mass $t \mapsto \rho_t(B_{\min\{R, c^2\eta^2\varepsilon/(8dL)\}}^\infty(x^*))$ is minimized at time $t = 0$. The latter assumption is motivated by numerical observations of typical successful CBO runs, where the particle density around the global minimizer tends to be minimized initially and steadily increases over time.

We notice the dependency of α_0 in (3.104) on the ambient dimension d , if we do not impose any additional structural assumption on \mathcal{E} , which might allow to replace the ambient dimension d by some notion of intrinsic dimensionality $d_{\text{intrinsic}} \ll d$.

Remark 3.21 (Computational complexity of CBO, cf. [CBO-I, Remark 14]). To achieve an accuracy of $\varepsilon_{\text{total}}$ as in Estimate (3.97) with probability of at least $(1 - 2\delta)$, the implementable CBO scheme (2.2) has to be run using $N \geq 9C_{\text{MFA}}/(\delta\varepsilon_{\text{total}})$ agents and with time step size $\Delta t \leq \sqrt[2m]{\delta\varepsilon_{\text{total}}/(18C_{\text{NA}})}$ for

$$K \geq \frac{1}{(1 - \vartheta)(2\lambda - \sigma^2)} \frac{1}{\Delta t} \log \left(\frac{36\mathcal{V}(\rho_0)}{\delta\varepsilon_{\text{total}}} \right) \quad (3.105)$$

iterations. Here, the parameter dependence of C_{NA} and C_{MFA} is as described in [Theorem 3.19](#). The computational complexity (counted in terms of the number of evaluations of the objective \mathcal{E}) of the CBO method is therefore given by $\mathcal{O}(KN)$.

When working in the setting of large-scale applications arising, for instance, in machine learning and signal processing (therefore, with \mathcal{E} being expensive to compute), several considerations allow to reduce the overall runtime of the algorithm [\(2.2\)](#) and thereby make the method feasible and more competitive. First of all, it may be recommendable to leverage that the evaluations of the objective function \mathcal{E} for each of the N particles can be performed in parallel. Furthermore, random mini-batch sampling ideas as proposed in [[Car+21](#); [CBO-II](#)] may be employed when evaluating the objective function and/or computing the consensus point. I.e., at each time step, \mathcal{E} is evaluated only on a random subset of the available data, and $x_\alpha^\mathcal{E}$ is computed only from a subset of the N particles. Besides immediately reducing the computational and communication complexity of CBO methods, such ideas motivate communication-efficient parallelization of the algorithm by evolving disjoint subsets of particles independently for some time with separate consensus points, before aligning the dynamics through a global communication step. This, however, is so far largely unexplored, both from a theoretical and practical point of view. Lastly, taking inspiration from genetic algorithms, a variance-based particle reduction technique as suggested in [[For+21](#)] may be used to reduce the number of optimizing agents (and therefore the required evaluations of \mathcal{E}) during the algorithm in case concentration of the particles is observed.

For a more in-depth discussion of topics related to the computational complexity of CBO and, in particular, on how to reduce the computational cost of CBO, we refer, amongst others, to [[CBO-I](#), Remark 14], [[CBO-II](#), Section 4], [[Car+21](#), Sections 2 and 4], and [[For+21](#), Section 2]. For implementational aspects, we refer to [[CBX](#)].

3.2. Global Convergence of Consensus-Based Optimization with Truncated Noise¹⁸

Computations of higher-order moments, conducted analogously to the ones required for the proof of [Lemma 3.15](#), suggest that these moments of the standard CBO dynamics [\(2.2\)](#) as well as of its continuous-time form [\(3.1\)](#) and mean-field limit [\(3.2\)](#) are not well-behaved and might exhibit characteristics of heavy tails. In order to enhance the well-behavedness of the statistics of the law of the dynamics, we explore in [[CBO-III](#)] a variant of CBO, which incorporates truncated noise.

Given a finite time horizon $T > 0$ and a time discretization $0 < \Delta t < \dots < K\Delta t = T$ of $[0, T]$ with a suitable discrete time step size $\Delta t > 0$, we denote the position of the i th agent at time step k again¹⁹ by $X_k^i \in \mathbb{R}^d$ and the empirical measure of all agents at time step k by $\widehat{\rho}_k^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}$. For user-specified parameters $\alpha, \lambda, \sigma > 0$ as well

¹⁸In this section, we follow [[CBO-III](#)].

¹⁹Throughout this section, but limited to it, we denote by $((X_k^i)_{k=1, \dots, K})_{i=1, \dots, N}$, $((X_t^i)_{t \geq 0})_{i=1, \dots, N}$, $(\bar{X}_t)_{t \geq 0}$, $(\widehat{\rho}_k^N)_{k=1, \dots, K}$, and $(\rho_t)_{t \geq 0}$ the quantities of CBO with truncated noise [\(3.106\)](#), [\(3.108\)](#), and [\(3.109\)](#) instead of standard CBO [\(2.2\)](#), [\(3.1\)](#), and [\(3.2\)](#).

as $x_b \in \mathbb{R}^d$ and $R, M > 0$, the time-discrete evolution of the i th particle in CBO with truncated noise is given by the iterative update rule

$$X_k^i = X_{k-1}^i - \Delta t \lambda \left(X_{k-1}^i - \mathbf{P}_{x_b, R} \left(x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^N) \right) \right) + \sigma \left(D \left(X_{k-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^N) \right) \wedge M \right) B_k^i, \quad (3.106)$$

where $\mathbf{P}_{x_b, R}$ is the projection onto $B_R(x_b)$ defined as

$$\mathbf{P}_{x_b, R}(x) := \begin{cases} x, & \text{if } \|x - x_b\|_2 \leq R, \\ x_b + R \frac{x - x_b}{\|x - x_b\|_2}, & \text{if } \|x - x_b\|_2 > R. \end{cases} \quad (3.107)$$

As a crucial assumption in this section, we assume that the map $\mathbf{P}_{x_b, R}$ depends on R and x_b in such way that $x^* \in B_R(x_b)$. Moreover, D is as described in (2.5) and, in the anisotropic case, \wedge acts elementwise and takes the absolute value of the left-hand side. The associated continuous-time analog reads

$$dX_t^i = -\lambda \left(X_t^i - \mathbf{P}_{x_b, R} \left(x_\alpha^\mathcal{E}(\widehat{\rho}_t^N) \right) \right) dt + \sigma \left(D \left(X_t^i - x_\alpha^\mathcal{E}(\widehat{\rho}_t^N) \right) \wedge M \right) dB_t^i. \quad (3.108)$$

For the global convergence analysis of (3.106) and (3.108), respectively, we follow the framework of [CBO-I; CBO-II], which we outlined in detail in the preceding Section 3.1. In particular, convergence to a global minimizer is first studied from a mean-field perspective, i.e., by analyzing the mono-particle process

$$d\bar{X}_t = -\lambda \left(\bar{X}_t - \mathbf{P}_{x_b, R} \left(x_\alpha^\mathcal{E}(\rho_t) \right) \right) dt + \sigma \left(D \left(\bar{X}_t - x_\alpha^\mathcal{E}(\rho_t) \right) \wedge M \right) dB_t. \quad (3.109)$$

By introducing this additional truncations in the CBO dynamics, we achieve that, in contrast to the original version, higher-order moments of the law of the dynamics can be effectively bounded. For an intuitive sketch highlighting the effects of the truncation, we refer to [CBO-III, Section 1]. More formally, however, it holds the following key result of [CBO-III].

Lemma 3.22 (Sub-Gaussianity of the mean-field dynamics, [CBO-III, Lemma 8]). Let R and M be finite such that $R \geq \|x_b - x^*\|_2$. For any $\psi > 0$, let N satisfy $N \geq (4\sigma^2 M^2)/(\lambda\psi^2)$. Moreover, let $((\bar{X}_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote N independent copies of the strong solution to the mean-field dynamics (3.109). Then, provided that $\mathbb{E} \exp \left(\sum_{i=1}^N \|\bar{X}_0^i - x^*\|_2^2 / (N\psi^2) \right) < \infty$, it holds

$$C_\psi := \sup_{t \in [0, T]} \mathbb{E} \exp \left(\frac{1}{N\psi^2} \sum_{i=1}^N \|\bar{X}_t^i - x^*\|_2^2 \right) < \infty, \quad (3.110)$$

where $C_\psi = C_\psi(\psi, \lambda, \sigma, d, R, M, T)$.

The proof of Lemma 3.22 is presented in [CBO-III, Section 3.2.1].

The sub-Gaussianity of \bar{X}_t follows from Lemma 3.22 by noticing that the statement can be applied in the setting $N = 1$ when choosing ψ sufficiently large.

As a consequence thereof and constituting the central contribution of [CBO-III], this variant exhibits enhanced convergence performance. On the one side, this is reflected from a practical point of view in [CBO-III, Figure 1, Section 4] and from a theoretical one in Theorem 3.25 by allowing for a wider flexibility in choosing the noise parameter of the method. In the case of anisotropic noise, instead of having the requirement $2\lambda > \sigma^2$, we need $\lambda \geq 2\sigma^2$ or $\sigma^2 M^2 = \mathcal{O}(\epsilon)$, where the latter allows for a trade-off between σ and M . On the other side, and more significantly, when adopting the analytical framework of [CBO-I] the gained regularity allows to establish a non-probabilistic mean-field approximation, see Proposition 3.24. This, in turn, enables us to prove global convergence in expectation rather than probability for the proposed CBO variant requiring only minimal assumptions on the objective function and on the initialization, see Theorem 3.25.

Let us now present the main result about global convergence of CBO with truncated noise for objective functions satisfying in addition to Assumption 3.5 the following.

Assumption 3.23. In this section, we are interested in objectives $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$, for which additionally

T1 there exist constants $\underline{\alpha}, L_u > 0$ such that for any $\alpha \geq \underline{\alpha}$ it holds

$$L_u := \sup_{x \in \mathbb{R}^d} \|x \exp(-\alpha(\mathcal{E}(x) - \underline{\mathcal{E}}))\|_2 < \infty, \quad (3.111)$$

T2 there exist constants $\gamma \in [0, 1]$ and $\tilde{C}_1, \tilde{C}_2 > 0$ such that

$$|\mathcal{E}(x) - \mathcal{E}(x')| \leq \tilde{C}_1 (\|x - x^*\|_2^\gamma + \|x' - x^*\|_2^\gamma) \|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d, \quad (3.112)$$

$$\mathcal{E}(x) - \underline{\mathcal{E}} \leq \tilde{C}_2 (1 + \|x - x^*\|_2^{1+\gamma}), \quad \text{for all } x \in \mathbb{R}^d. \quad (3.113)$$

T1 requires a certain growth of the function \mathcal{E} . T2 sets controllable bounds on the local Lipschitz constant of \mathcal{E} and on the growth of \mathcal{E} , which is required to be at most quadratic. A similar requirement appears also in Assumption 3.3, but a quadratic lower bound was also imposed.

Under these assumptions, we first have the formerly addressed non-probabilistic mean-field approximation result, which improves [CBO-I, Proposition 16] and Proposition 3.16 by being non-probabilistic. Of course, let us emphasize that this required the modification of the standard CBO dynamics by introducing truncated noise.

Proposition 3.24 (Mean-field approximation of CBO with truncated noise,

cf. [CBO-III, Proposition 7]). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2 as well as T1–T2. Moreover, let R and M be finite such that $R \geq \|x_b - x^*\|_2$ and let $N \geq (16\alpha\tilde{C}_2\sigma^2M^2)/\lambda$. Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N \in \mathbb{N}$ be fixed. Moreover, let $((X_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote the strong solution to system (3.108) and let $((\bar{X}_t^i)_{t \geq 0})_{i=1, \dots, N}$ be N independent copies of the strong solution to the mean-field dynamics (3.109). If $(X_t^i)_{t \geq 0}$ and $(\bar{X}_t^i)_{t \geq 0}$ share the initial data as well as the Brownian motion paths $(B_t^i)_{t \geq 0}$ for all $i = 1, \dots, N$, then

we have

$$\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \left(X_T^i - \bar{X}_T^i \right) \right\|_2^2 \leq C_{\text{MFA}} N^{-1} \quad (3.114)$$

with $C_{\text{MFA}} = C_{\text{MFA}}(\alpha, \lambda, \sigma, d, T, \nu, \eta, \tilde{C}_1, \tilde{C}_2, L_u, R, x_b, x^*, M)$.

The proof of [Proposition 3.24](#) is presented in [[CBO-III](#), Section 3.2.1].

Combining this statement with a convergence result about the mean-field dynamics (3.109) as derived in [[CBO-III](#), Section 3.2.2], yields the main result of [[CBO-III](#)] about global convergence of CBO with truncated noise in expectation. Notice that convergence in expectation is stronger than convergence in probability, i.e., implies the latter.

Theorem 3.25 (CBO with truncated noise converges globally, cf. [[CBO-III](#), Theorem 3]). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy [A1–A2](#) as well as [T1–T2](#). Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ be such that $x^* \in \text{supp}(\rho_0)$. Define $\mathcal{V}(\rho_t)$ as given in (3.18). Provided that $\mathcal{V}(\rho_0) > 0$, fix any $\varepsilon \in (0, \mathcal{V}(\rho_0))$, let $R \in (\|x_b - x^*\|_2 + \sqrt{\varepsilon/2}, \infty)$ and $M \in (0, \infty)$. Choose parameters $\lambda, \sigma > 0$ with $\lambda \geq 2\sigma^2$ or $\sigma^2 M^2 = \mathcal{O}(\varepsilon)$, and define the time horizon

$$T^* := \frac{1}{\lambda} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right). \quad (3.115)$$

Then with $K := T^*/\Delta t$ for suitable $\Delta t > 0$ and by choosing α sufficiently large and $N \geq (16\alpha\tilde{C}_2\sigma^2 M^2)/\lambda$, the iterations $((X_k^i)_{k=1,\dots,K})_{i=1,\dots,N}$ generated by the numerical scheme (3.106) converge in expectation to x^* . More precisely, the empirical mean of the final iterations fulfills

$$\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N X_K^i - x^* \right\|_2^2 \lesssim C_{\text{NA}} (\Delta t)^{2m} + C_{\text{MFA}} N^{-1} + \varepsilon \quad (3.116)$$

up to a generic constant. Here, m denotes the order of accuracy of the numerical scheme (for the Euler-Maruyama scheme $m = 1/2$). Moreover, besides problem-dependent constants and parameters of the method, $C_{\text{NA}} > 0$ depends linearly on the dimension d and the number of particles N , and exponentially on the time horizon T^* ; $C_{\text{MFA}} > 0$ depends exponentially on the parameters α, λ and σ , and on T^* .

The proof of [Theorem 3.25](#) is presented in [[CBO-III](#), Section 3.2].

3.3. Global Convergence of Consensus-Based Optimization with Memory Effects and Gradient Information²⁰

Taking inspiration from the typically more intricate dynamics of interacting multi-particle systems employed for optimization in practical applications, we wrap up this

²⁰In this section, we follow [[CBO-IV](#)].

chapter with a more elaborate variant of the standard CBO dynamics (2.2) proposed and analyzed in [CBO-IV]. It exhibits two additional features, namely memory effects and gradient information. While the purpose of the latter is self-explanatory by permitting each particle of the algorithm to exploit local information about the energy landscape through gradients, the former mechanisms equip the particles with a memory of their historical positions. More precisely, realizing their implementation as suggested in the work [GP21], we introduce²¹ for each particle X^i an additional state variable Y^i , which stores the historical best position of the respective particle X^i . Consequently, an individual particle is described by the tuple (X^i, Y^i) . An alternative realization of memory mechanisms, which, however, might require substantially different analysis techniques, is proposed by the authors of [TW20]. Both additional information, memory effects as well as gradient information, are exploited in the CBO dynamics through drift terms, i.e., in addition to the standard consensus drift, each particle experiences a drift to its personal historical best position as well as in the direction of the local gradient. To further enhance the exploration capabilities of the method and to allow for a mathematical analysis, the new drift terms are also accompanied by associated noise terms. Moreover, the consensus point is no longer computed from the instantaneous positions X^i , but the historical best positions Y^i . We, therefore, denote it by $y_\alpha^\mathcal{E}$. Let us now make the description of this variant more rigorous.

Given a finite time horizon $T > 0$ and a time discretization $0 < \Delta t < \dots < K\Delta t = T$ of $[0, T]$ with a suitable discrete time step size $\Delta t > 0$, we denote the position of the i th agent at time step k by $X_k^i \in \mathbb{R}^d$, its historical best position stored in the particle's memory as $Y_k^i \in \mathbb{R}^d$, and the empirical measure of all agents' historical best positions at time step k by $\hat{\rho}_{Y,k}^N := \frac{1}{N} \sum_{i=1}^N \delta_{Y_k^i}$. For user-specified parameters $\alpha, \beta, \theta, \kappa, \lambda_1, \sigma_1 > 0$ and $\lambda_2, \lambda_3, \sigma_2, \sigma_3 \geq 0$, the time-discrete evolution of the i th particle in CBO with memory effects and gradient information is given by the iterative update rule

$$\begin{aligned} X_k^i &= X_{k-1}^i - \Delta t \lambda_1 \left(X_{k-1}^i - y_\alpha^\mathcal{E}(\hat{\rho}_{Y,k-1}^N) \right) + \sigma_1 D \left(X_{k-1}^i - y_\alpha^\mathcal{E}(\hat{\rho}_{Y,k-1}^N) \right) B_k^{1,i} \\ &\quad - \Delta t \lambda_2 \left(X_{k-1}^i - Y_{k-1}^i \right) + \sigma_2 D \left(X_{k-1}^i - Y_{k-1}^i \right) B_k^{2,i} \\ &\quad - \Delta t \lambda_3 \nabla \mathcal{E}(X_{k-1}^i) + \sigma_3 D \left(\nabla \mathcal{E}(X_{k-1}^i) \right) B_k^{3,i}, \end{aligned} \quad (3.117a)$$

$$Y_k^i = Y_{k-1}^i + \Delta t \kappa \left(X_k^i - Y_{k-1}^i \right) S^{\beta, \theta} \left(X_k^i, Y_{k-1}^i \right), \quad (3.117b)$$

where $((B_k^{m,i})_{k=1,\dots,K})_{i=1,\dots,N}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}$ for $m \in \{1, 2, 3\}$. The system is complemented with independent initial data $(X_0^i, Y_0^i)_{i=1,\dots,N}$, typically such that $X_0^i = Y_0^i$ for all $i = 1, \dots, N$. In addition to the terms familiar from standard CBO in the first line of (3.117a), the first term in the second line of (3.117a) is the drift towards the historical best position of the respective particle. In contrast to the global nature of the consensus drift, which incorporates information from all N particles, this term

²¹Throughout this section, but limited to it, we denote by $((X_k^i)_{k=1,\dots,K})_{i=1,\dots,N}$, and $(\rho_t)_{t \geq 0}$ the quantities of CBO with memory effects and gradient information (3.117a), and (3.119) instead of standard CBO (2.2), and (3.3).

depends only on the past of the specific particle i . To store such information about the history of each particle [GP21], the additional state variable Y^i evolves according to (3.117b), where

$$S^{\beta,\theta}(x, y) = \frac{1}{2}(1 + \theta + \tanh(\beta(\mathcal{E}(y) - \mathcal{E}(x)))) \quad (3.118)$$

is chosen to approximate the Heaviside function $H(x, y) = \mathbb{1}_{\mathcal{E}(x) < \mathcal{E}(y)}$ as $\theta \rightarrow 0$ and $\beta \rightarrow \infty$. Y_k^i can therefore be regarded as the memory of the i th particle, i.e., as the location of the in-time best-seen position of X^i up to time step k . This can be understood when noticing that with parameter choices $\kappa = 1/\Delta t$, $\theta = 0$ and $\beta \gg 1$ in (3.117b) it holds $Y_k^i = X_k^i$ if $\mathcal{E}(X_k^i) < \mathcal{E}(Y_{k-1}^i)$ and $Y_k^i = Y_{k-1}^i$ else. The first term in the third line of (3.117a) is the drift in the direction of the negative gradient of \mathcal{E} , which is a local and instantaneous contribution. Eventually, the remaining two terms are noise terms, which are associated with the formerly described memory and gradient drifts.

The central theoretical contribution of [CBO-IV] is the global convergence analysis of (3.117) from a mean-field perspective following the framework put forward in [CBO-I; CBO-II] and as done in Section 3.1.1 for the standard CBO dynamics (2.2). In analogy to the derivations there, we find that the macroscopic continuous-time description of (3.117) is given by the measure $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ which satisfies the nonlinear nonlocal Fokker-Planck equation

$$\begin{aligned} \partial_t \rho_t = & \operatorname{div}_x \left(\left(\lambda_1 (x - y_{\alpha}^{\mathcal{E}}(\rho_{Y,t})) + \lambda_2 (x - y) + \lambda_3 \nabla \mathcal{E}(x) \right) \rho_t \right) \\ & + \operatorname{div}_y \left(\left(\kappa (y - x) S^{\beta,\theta}(x, y) \right) \rho_t \right) \\ & + \frac{1}{2} \sum_{k=1}^d \partial_{x_k x_k}^2 \left(\left(\sigma_1^2 D(x - y_{\alpha}^{\mathcal{E}}(\rho_{Y,t}))_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(x))_{kk}^2 \right) \rho_t \right) \end{aligned} \quad (3.119)$$

in a weak sense (see [CBO-IV, Definition 1]). Analyzing (3.119) in place of (3.117) typically permits to employ more powerful technical tools, which result in stronger statements about the long-time behavior of the average agent density ρ . This analysis approach is rigorously justified by the mean-field approximation, i.e., the fact that the empirical particle measure $\widehat{\rho}_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^i, Y_t^i)}$ converges in some sense to the mean-field law ρ_t as the number of particles N tends to infinity, see Section 3.1.2 for more information. While a quantitative result about the mean-field approximation of this variant is left for future considerations, qualitatively, the convergence can be shown by following [Hua21], see in particular [Hua21, Remark 3.2].

This justifies the analysis of the dynamics on the macroscopic level (3.119) to gain insights into its behavior. Let us, therefore, present the main result about global convergence of CBO with memory effects and gradient information in mean-field law for objective functions satisfying in addition to Assumption 3.5 the following.

Assumption 3.26. In this section and for the case of an additional gradient drift component, i.e., if $\lambda_3 \neq 0$, we additionally require that $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ and that

G1 there exist $C_{\nabla\mathcal{E}} > 0$ such that

$$\|\nabla\mathcal{E}(x)\|_2 \leq C_{\nabla\mathcal{E}} \|x - x^*\|_2, \quad \text{for all } x \in \mathbb{R}^d. \quad (3.120)$$

In case of an additional gradient drift term in the dynamics, i.e., $\lambda_3 \neq 0$, the objective naturally needs to be continuously differentiable. Furthermore, G1 imposes that the gradient $\nabla\mathcal{E}$ grows at most linearly. This is a significantly weaker assumption compared to typical smoothness assumptions about \mathcal{E} in the optimization literature (in particular in the analysis of stochastic gradient descent), where Lipschitz-continuity of the gradient of \mathcal{E} is required [MB11].

Under these assumptions, we have the following statement about global convergence of CBO with memory effects and gradient information in mean-field law.

Theorem 3.27 (CBO with memory and gradient converges globally in mean-field law, [CBO-IV, Theorem 2.5]). Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Furthermore, in the case of an active gradient drift, i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ obey in addition G1. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$ be such that $(x^*, x^*) \in \text{supp}(\rho_0)$. Define the functional

$$\mathcal{V}(\rho_t) := \frac{1}{2} \iint (\|x - x^*\|_2^2 + \|y - x^*\|_2^2) d\rho_t(x, y), \quad (3.121)$$

and the rates

$$\chi_1 := \min \left\{ \lambda_1 - \lambda_2 - 3\lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_1^2 - 2\sigma_3^2 C_{\nabla\mathcal{E}}^2, 2\kappa\theta + \lambda_2 - \lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_2^2 \right\}, \quad (3.122a)$$

$$\chi_2 := \max \left\{ 3\lambda_1 + \lambda_2 + 3\lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_1^2 + 2\sigma_3^2 C_{\nabla\mathcal{E}}^2, 2\kappa\theta + 3\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_2^2 \right\}, \quad (3.122b)$$

which we assume to be strictly positive through a sufficient choice of the parameters of the CBO dynamics. Provided that $\mathcal{V}(\rho_0) > 0$, fix any $\varepsilon \in (0, \mathcal{V}(\rho_0))$, $\vartheta \in (0, 1)$ and define the time horizon

$$T^* := \frac{1}{(1 - \vartheta)\chi_1} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right). \quad (3.123)$$

Then there exists $\alpha_0 > 0$, depending (among problem dependent quantities) also on ε and ϑ , such that for all $\alpha > \alpha_0$, if $\rho \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (3.119) on the time interval $[0, T^*]$ with initial condition ρ_0 , we have

$$\mathcal{V}(\rho_T) = \varepsilon \quad \text{with} \quad T \in \left[\frac{(1 - \vartheta)\chi_1}{(1 + \vartheta/2)\chi_2} T^*, T^* \right]. \quad (3.124)$$

Furthermore, on the time interval $[0, T]$, $\mathcal{V}(\rho_t)$ decays at least exponentially fast. More precisely, for all $t \in [0, T]$ it holds

$$W_2^2(\rho_t, \delta_{(x^*, x^*)}) \leq 6\mathcal{V}(\rho_t) \leq 6\mathcal{V}(\rho_0) \exp(-(1 - \vartheta)\chi_1 t). \quad (3.125)$$

The proof of [Theorem 3.27](#) presented in [[CBO-IV](#), Section 3] reveals how to leverage further, in other applications advantageous, forces in the dynamics while still being amenable to theory and allowing for provable global convergence within the framework of [[CBO-I](#); [CBO-II](#)].

The benefit of the herein investigated CBO variant exploiting memory effects and gradient information is demonstrated in [[CBO-IV](#), Figure 1, Section 4] for a benchmark problem in optimization as well as for certain applications coming from machine learning and signal processing.

Chapter 4

Interpreting Consensus-Based Optimization as a Stochastic Relaxation of Gradient Descent

Leaving aside for the moment the mean-field analysis perspective taken in the rest of this dissertation, we discuss in this chapter the second core insight about the optimization behavior of CBO addressed in this thesis. It is concerned with the observation that, despite solely relying on evaluations of the objective, through communication of the particles, CBO exhibits a stochastic gradient descent (GD)-like behavior. This is revealed when studying the trajectory of the consensus point of the method and sheds light on the CBO algorithm (2.2) from a different angle. In Section 4.1 we give an overview of the main contributions of [CBO&GD], which are about the interpretation of CBO as a stochastic relaxation of GD with a problem-tailored stochastic perturbation. The theoretical results are corroborated by instructive numerical illustrations. The proof idea of establishing such a bridge between a metaheuristic black-box and derivative-free optimization algorithms on the one hand and a gradient-based learning method on the other is sketched in Section 4.2, where we in particular introduce the consensus hopping (CH) scheme, which connects CBO with GD. With the results of Chapter 3 about the provable global convergence capabilities of CBO in mind, the fundamental value of such link between CBO and stochastic GD lies in offering, on the one side, a novel explanation for the success of stochastic relaxations of GD and providing a novel point of view on the theoretical understanding of gradient-based learning algorithms, while, on the other side, unveiling an intrinsic GD nature of such heuristics.

4.1. Consensus-Based Optimization Exhibits a Stochastic Gradient Descent-Like Behavior²²

An insightful theoretical understanding of the behavior of CBO methods is to be gained by tracing the dynamics of the consensus point $x_\alpha^\mathcal{E}$ of the CBO algorithm (2.2). To this end, let us introduce the CBO scheme as the iterates $(x_k^{\text{CBO}})_{k=0,\dots,K}$ defined according

²²In this section, we follow [CBO&GD, Sections 1 and 4].

to

$$\begin{aligned} x_k^{\text{CBO}} &= x_\alpha^\mathcal{E}(\widehat{\rho}_k^N), \quad \text{with} \quad \widehat{\rho}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}, \\ x_0^{\text{CBO}} &= x_0 \sim \rho_0, \end{aligned} \tag{4.1}$$

where we recall that the positions X_k^i of the particles are given by the iterative update rule (2.2).

Constituting the main contribution of the work [CBO&GD] is the novel observation as well as theoretical and experimental justification that the iterates of the CBO scheme (4.1), i.e., the trajectory of the consensus point $x_\alpha^\mathcal{E}$, follow, with high probability, the path of a stochastically perturbed GD with the stochastic perturbation being problem-tailored. We make this rigorous in Theorem 4.2 and demonstrate it numerically in Figure 4.1 below.

To the best of our knowledge, this is the first attempt of its kind to interconnect the derivative-free with the gradient-based world in optimization. An in spirit similar observation has been made recently in [Par24], where the Langevin dynamics [GH86; CHS87] is recovered in a suitable scaling limit from simulated annealing [KGV83; Kir84; AK89] by using tools from linear kinetic theory.

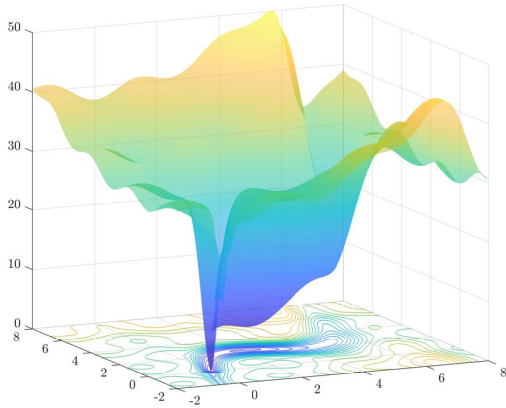
Let us now present the main findings of [CBO&GD]. The class of objective functions considered in our theoretical results below satisfies in addition to Assumption 3.3 and A1 from Assumption 3.5 the following.

Assumption 4.1. In this chapter, we additionally require that $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ and that the objective is

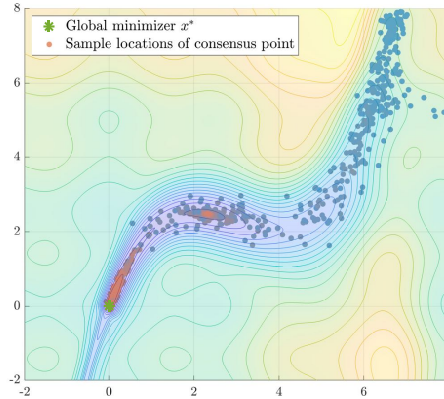
- G1 semi-convex (Λ -convex for some $\Lambda \in \mathbb{R}$), i.e., $\mathcal{E}(\bullet) - \frac{\Lambda}{2} \|\bullet\|_2^2$ is convex,
- G2 L -smooth, i.e.,

$$\|\nabla \mathcal{E}(x) - \nabla \mathcal{E}(x')\|_2 \leq L \|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d. \tag{4.2}$$

G1 requires the objective \mathcal{E} to be semi-convex with parameter $\Lambda \in \mathbb{R}$. For $\Lambda > 0$, Λ -convexity is stronger than convexity (strong convexity with parameter Λ). For $\Lambda < 0$, semi-convexity is weaker, i.e., potentially nonconvex functions \mathcal{E} are included in the definition. The class of semi-convex functions is typical in the literature of gradient flows, since their general theory extends from the convex to this more general setting [San17]. One particular property, which we shall exploit in this work, is that for such functions the time discretization of a gradient flow, potentially for a small step size, defined through an iterated scheme, called minimizing movement scheme [De 93], is well-defined. However, while semi-convexity is useful to ensure the well-posedness of gradient flows, it is not sufficient to obtain convergence to global minimizers. Other properties such as the Polyak-Łojasiewicz (PL) condition [KNS16], which requires that for some $\mu > 0$ it holds $\|\nabla \mathcal{E}(x)\|_2^2 / 2 \geq \mu(\mathcal{E}(x) - \mathcal{E})$ for all $x \in \mathbb{R}^d$, or the log-Sobolev inequalities governing the flow of the Langevin dynamics [CB18] may be necessary. Concerning the PL condition,



(a) A noisy Canyon function with a valley shaped as a third degree polynomial in two dimensions as objective function \mathcal{E} .



(b) CBO can be interpreted as a stochastic relaxation of GD.

The CBO scheme (4.1) (sampled over several runs) follows on average the valley of \mathcal{E} while passing over local minima.

Figure 4.1: An illustration of the intuition that the CBO scheme (4.1) can be regarded as a stochastic derivative-free (zero-order) relaxation of GD. To find the global minimizer x^* (green star) of the nonconvex objective function \mathcal{E} depicted in (a), we run the CBO algorithm (2.2) for $K = 250$ iterations with parameters $\Delta t = 0.01$, $\alpha = 100$, $\lambda = 1$ and $\sigma = 1.6$, and $N = 200$ particles, initialized i.i.d. according to $\rho_0 = \mathcal{N}((8, 8), 0.5 \text{Id})$. This experiment is performed 50 times. For each run we depict in (b) the positions of the consensus points computed during the CBO algorithm (2.2), i.e., the iterates of the CBO scheme (4.1) for $k = 1, \dots, K$. The color of the individual points corresponds to time, i.e., iterates at the beginning of the scheme are plotted in blue, whereas later iterates are colored orange. We observe that, after starting close to the initial position, the trajectories of the consensus points follow the path of the valley leading to the global minimizer x^* , until it is reached. In particular, unlike GD (cf. [CBO&GD, Figure 2b]), the scheme (4.1) has the capability of jumping over locally deeper passages. Such desirable behavior is observed also for the Langevin dynamics (see Figure [CBO&GD, Figure 2c]), which can be regarded as a stochastic (noisy) version of GD.

notice that it does not imply that there is a unique solution, but it implies that every stationary point is a global minimum. G2 assumes smoothness of the objective function \mathcal{E} by requiring the gradient $\nabla \mathcal{E}$ to be L -Lipschitz continuous. In the realm of machine learning, in particular, when using gradient-based methods for optimization, this is a quite standard assumption, as it assures that the gradient information is informative within a region around where the gradient is evaluated. In particular, as proved in [Pol63], under the assumptions of L -smoothness and the PL-condition, one obtains global linear convergence of GD. More precisely, for the GD iteration $x_k = x_{k-1} + \frac{1}{L} \nabla \mathcal{E}(x_{k-1})$ it holds $\mathcal{E}(x_k) - \underline{\mathcal{E}} \leq (1 - \frac{\mu}{L})^k (\mathcal{E}(x_0) - \underline{\mathcal{E}})$, see [KNS16, Theorem 1].

Under these assumptions, we have the following statement.

Theorem 4.2 (CBO is a stochastic relaxation of GD, [CBO&GD, Theorem 1]). Let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ satisfy W1–W2, A1, and G1–G2. Then, for $\tau > 0$ (satisfying $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$) and with parameters $\alpha, \lambda, \sigma, \Delta t > 0$ such that $\alpha \gtrsim \frac{1}{\tau} d \log d$, the iterates $(x_k^{\text{CBO}})_{k=0, \dots, K}$ of the CBO scheme (4.1) follow a stochastically perturbed GD, i.e., they obey

$$x_k^{\text{CBO}} = x_{k-1}^{\text{CBO}} - \tau \nabla \mathcal{E}(x_{k-1}^{\text{CBO}}) + g_k, \quad (4.3)$$

where g_k is stochastic noise fulfilling, with high probability, the quantitative estimate

$$\|g_k\|_2 = \mathcal{O} \left(|\lambda - 1/\Delta t| + \sigma \sqrt{\Delta t} + \sqrt{\tau/\alpha} + N^{-1/2} \right) + \mathcal{O}(\tau) \quad (4.4)$$

for each $k = 1, \dots, K$.

The proof of Theorem 4.2 is presented in [CBO&GD, Section 4.1 as well as Section 4.2 and Appendices C–E].

Let us conclude this section about a review of the contributions of our work [CBO&GD] with some remarks regarding the interpretation of the result.

Remark 4.3 (Interpretation of Theorem 4.2, [CBO&GD, Discussion after Theorem 1]). The statement of Theorem 4.2 has to be read with a twofold interpretation. First, in view of the capability of CBO to converge to global minimizers for rich classes of nonsmooth and nonconvex objective functions (see [CBO-I, Theorem 13] and Theorem 3.19), Theorem 4.2 states that there exist stochastic relaxations of GD that are provably able to robustly and reliably overcome energy barriers and reach deep levels of nonconvex functions. Such relaxations may even be derivative-free and do not require smoothness of the objective, as is the case with CBO. Second, and conversely, against the common wisdom that derivative-free optimization heuristics search the domain mainly by random exploration and therefore ought to be inefficient, we provide evidence that such heuristics in fact work successfully in finding benign optima [Duc+15; NS17; Che+17; Nik+22; Chi+23; ERY24; HWO23], precisely because they can be interpreted as suitable stochastic relaxations of gradient-based methods.

The interpretation of the CBO scheme (4.1) as a stochastic relaxation of GD is substantiated visually, analytically and numerically as follows. While the trajectories of (4.1) are to be seen in Figure 4.1b, we depict for comparison in [CBO&GD, Figure 2c] the discretized dynamics of the annealed Langevin dynamics [CHS87; RT96; DM17],

$$dX_t = -\nabla \mathcal{E}(X_t) dt + \sqrt{2\beta_t^{-1}} dB_t. \quad (4.5)$$

Both stochastic methods are capable of global minimization while overcoming energy barriers and escaping local minima. For analyses of the (annealed) Langevin dynamics we refer the reader to [GM91; Már97; Pel98] as well as the more recent works [Xu+18; Chi22]. The stochastic perturbations g_k in (4.3) are meaningful and not generic as they obey precise scalings thanks to the established estimate in (4.4). In particular,

as reflected by the first term of the bound on the error $\|g_k\|_2$, they become tighter as soon as the discrete CBO time step size $\Delta t \ll 1$, the drift parameter $\lambda \approx 1/\Delta t$, the noise parameter σ becomes smaller, the weight parameter α is sufficiently large, and the number of employed particles N becomes larger. This behavior is confirmed numerically in [Figure 4.2](#) below by measuring the closeness between the trajectories of the CBO scheme (4.1) and GD. More precisely, better approximation is achieved for the values of λ closer to $1/\Delta t$ (compare lines with different colors but same line style, and notice that smaller error can be obtained for larger λ), larger choices of N (compare different line styles within a color), and σ as small as possible (each line decreases as σ decreases). For fixed λ and N , however, σ needs to be sufficiently large (in particular in case of a fixed number of iterates K) to allow the CBO scheme (4.1) to iteratively explore the energy landscape within the time horizon. As visible from [Figure 4.2](#), a larger number of particles N is needed to pass to smaller σ and thus better approximation. Regarding the second term of the bound on the error $\|g_k\|_2$, we conjecture a potential amelioration of the estimate by refining the quantitative Laplace principle, [Proposition 3.12](#) or [[CBO-I](#), [Proposition 21](#)], involved in the proof of [[CBO&GD](#), [Proposition 7](#)], which would allow to remove the order $\mathcal{O}(\tau)$ dependence of the bound. Yet, as it stands, this term is about a deterministic bounded perturbation of the gradient, which is possibly of smaller magnitude than the gradient. Such bounded perturbation alone does not allow to overcome local energy barriers in general (just think of a local minimizer, around which the magnitude of gradients grows faster than the displacement: any movement from the minimizer ought necessarily to get reverted). Hence, it is the stochastic part of the perturbation that enables the convergence to global minimizers. In fact, for a moderate time step size $\Delta t > 0$, a drift parameter $\lambda > 0$ relatively small compared to $1/\Delta t$, a non-insignificant noise parameter $\sigma > 0$, a moderate value of the weight parameter $\alpha > 0$ and a modest number N of particles, CBO is factually a stochastic relaxation of GD with strong noise.

Apart from gaining primarily theoretical insights from this link, let us conclude this remark by mentioning a further, more practical aspect of establishing such a connection. In several real-world applications, including various machine learning settings, using gradients may be undesirable or even not feasible. This can be due to the black-box nature or nonsmoothness of the objective, memory limitations constraining the use of automatic differentiation, a substantial presence of spurious local minima, or the fact that gradients carry relevant information about data, which one may wish to keep private. In machine learning, in specific, the problems of hyperparameter tuning [[Ber+11](#); [RT18](#)], convex bandits [[Aga+11](#); [Sha17](#)], reinforcement learning [[SB98](#)], the training of sparse and pruned neural networks [[Hoe+21](#)], and federated learning [[SS15](#); [McM+17](#)] stimulate interest in methods alternative to gradient-based ones. In such situations, if one still wishes to rely on a GD-like optimization behavior, [Theorem 4.2](#) suggests the use of CBO, which will be both reliable and efficient.²³

²³Needlessly to be said, but if gradients are available and cheap to compute, methods which exploit this information are expected to be more efficient and competitive. However, incorporating a gradient drift into CBO is possible and may bear advantages of theoretical and practical nature [[CBO-IV](#); [STW23](#); [Car+23](#)].

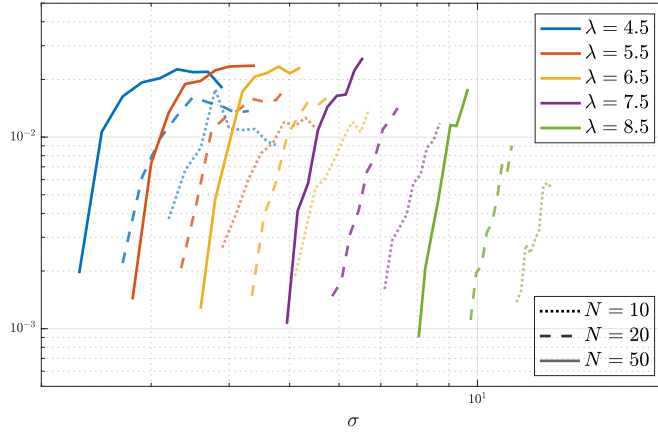


Figure 4.2: Numerical analysis of the approximation error between the trajectories of the CBO scheme (4.1) and GD, i.e., the stochastic noise g_k in (4.3). In the setting of the Canyon function \mathcal{E} from Figure 4.1a but without a local minimum in the valley,²⁴ we measure the distance between the two trajectories and plot the resulting approximation error for different values of λ (different colors), σ (horizontal axis), and N (different line styles). The other parameters of the CBO scheme (4.1) are $K = 1000$, $\Delta t = 0.1$ and $\alpha = 10^{16}$ with the remaining setting being as in Figure 4.1b. The results validate the theoretical bound on $\|g_k\|_2$ of Theorem 4.2.

4.2. From Consensus-Based Optimization to Consensus Hopping to Gradient Descent²⁵

In order to intuitively grasp how to establish a connection between the CBO scheme (4.1), which captures the flow of the derivative-free CBO dynamics (2.2), and GD, we provide in this section a brief overview of the proof idea for Theorem 4.2. To this end, the CH scheme is introduced in (4.6) below, which acts as the pillar of the bridge between CBO and GD.

It may moreover constitute a numerical method of independent interest, both from a practical and theoretical perspective. In fact, its resemblance to the covariance matrix adaptation evolution strategy (CMA-ES) [HO96; HO01; Han06; Oll+17] beckons further investigations. Particularly also in view of the latter being an instantiation of the natural evolution strategy (NES) [Wie+08; Wie+14; Sun+09], see, e.g., [Gla+10; Aki+10].

From CBO to CH. Let us envision the movement of the particles during the CBO dynamics (2.2). At every time step k , after having computed $x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)$, each particle moves a $\Delta t \lambda$ fraction of its distance towards this consensus point, before being perturbed by stochastic noise. As we let $\lambda \rightarrow 1/\Delta t$, the particles' velocities increase, until, in the case $\lambda = 1/\Delta t$, each of them hops within the Δt -time window directly to the previously

²⁴Otherwise, GD will necessarily get stuck in this local minimum located in the valley.

²⁵In this section, we follow [CBO&GD, Section 4].

computed consensus point, followed by a random fluctuation. Put differently, we are left with a numerical scheme, which, at time step k , samples N particles around the old iterate in order to subsequently compute as new iterate the consensus point (2.3) of the empirical measure of the samples. Such algorithm is precisely a Monte Carlo approximation of the CH scheme with iterates $(x_k^{\text{CH}})_{k=0,\dots,K}$ defined by

$$\begin{aligned} x_k^{\text{CH}} &= x_\alpha^\mathcal{E}(\mu_k), \quad \text{with} \quad \mu_k = \mathcal{N}\left(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id}\right), \\ x_0^{\text{CH}} &= x_0. \end{aligned} \tag{4.6}$$

In words, at time step $k = 1, \dots, K$, the new CH iterate is computed as the consensus point (2.3) w.r.t. a Gaussian distribution with covariance matrix $\tilde{\sigma}^2 \text{Id}$ centered at the old iterate. [CBO&GD, Theorem 6] makes this intuition rigorous by quantifying the approximation quality between the CBO and the CH scheme in terms of the parameters of the two schemes. Sample trajectories of the CH scheme are depicted in [CBO&GD, Figure 2a].

From CH to GD. With the sampling measure μ_k assigning (in particular for small $\tilde{\sigma}$) most mass to the region close to the old iterate, the CH scheme (4.6) improves at every time step k its objective function value while staying near the previous iterate. A conceptually analogous behavior to such localized sampling can be achieved through penalizing the length of the step taken at time step k . This gives rise to an implicit version of the CH scheme with iterates $(\tilde{x}_k^{\text{CH}})_{k=0,\dots,K}$ given as

$$\begin{aligned} \tilde{x}_k^{\text{CH}} &= \arg \min_{x \in \mathbb{R}^d} \tilde{\mathcal{E}}_k(x), \quad \text{with} \quad \tilde{\mathcal{E}}_k(x) := \frac{1}{2\tau} \left\| x_{k-1}^{\text{CH}} - x \right\|_2^2 + \mathcal{E}(x), \\ \tilde{x}_0^{\text{CH}} &= x_0. \end{aligned} \tag{4.7}$$

Actually, the modulated objective $\tilde{\mathcal{E}}_k$ defined in (4.7) naturally appears when writing out the expression of $x_\alpha^\mathcal{E}(\mu_k)$ from (4.6) using that μ_k is a Gaussian. This creates a link between the sampling width $\tilde{\sigma}$ and the step size τ . The fact that the parameter τ can be seen as the step size of (4.7) becomes apparent when observing that the optimality condition of the k -th iterate of (4.7) reads $\tilde{x}_k^{\text{CH}} = x_{k-1}^{\text{CH}} - \tau \nabla \mathcal{E}(\tilde{x}_k^{\text{CH}})$, which is an implicit gradient step. [CBO&GD, Proposition 7] estimates the discrepancy between x_k^{CH} and \tilde{x}_k^{CH} employing the quantitative Laplace principle, see Proposition 3.12 or [CBO-I, Proposition 21].

Let us conclude this discussion by remarking that the scheme (4.7) itself is not self-consistent but requires the computation of the iterates of the CH scheme (4.6). For this reason, we introduce the minimizing movement scheme (MMS) [De 93] as the iterates $(x_k^{\text{MMS}})_{k=0,\dots,K}$ given according to

$$\begin{aligned} x_k^{\text{MMS}} &= \arg \min_{x \in \mathbb{R}^d} \mathcal{E}_k(x), \quad \text{with} \quad \mathcal{E}_k(x) := \frac{1}{2\tau} \left\| x_{k-1}^{\text{MMS}} - x \right\|_2^2 + \mathcal{E}(x), \\ x_0^{\text{MMS}} &= x_0, \end{aligned} \tag{4.8}$$

which is known to be the discrete-time implicit Euler of the gradient flow dynamics $\frac{d}{dt}x(t) = -\nabla \mathcal{E}(x(t))$, see, e.g., [San17].

Chapter 5

Particle Swarm Optimization

The next core contribution of this dissertation, which we present in this chapter, is concerned with a convergence analysis of the renowned and widely-used PSO method [KE95; Ken97]. The method is popular among practitioners and recognized as an efficient and practicable black box algorithm for tackling complex optimization problems of the form (2.1) [WTL18]. As sketched at the end of Section 2.1 by following [GP21], PSO was not only the source of inspiration for the design of CBO but is rigorously related to the CBO methods discussed so far and can be regarded as a second-order variant of CBO that includes velocity and momentum [CHQ22]. In Section 5.1, we briefly recall the formulation of PSO used for our convergence analysis, which is then the focus of Section 5.2, where we give an overview of the theoretical results of [PSO]. By transferring the lessons and techniques learned from CBO to PSO, we investigate the convergence behavior of PSO to global minimizers under certain conditions of well-preparation of the hyperparameters of the method and the initial datum. This analysis employs the technique of [Car+18; Car+21]. An analysis in the framework of [CBO-I; CBO-II; CBO-IV] is subject of ongoing work.

5.1. The Dynamics of Particle Swarm Optimization²⁶

As alluded to at the end of Section 2.1 as well as at the beginning of Section 3.3, the dynamics of CBO given in (2.2) is, as intended by the authors of [Pin+17], kept simple and idealized compared to the ones of classical PSO methods [KE95; Ken97; SE98] or other related particle-based optimization algorithms [DB05; Pha+06; Fil+08]. The intricate working principles of these methods, namely, impede a rigorous study of their convergence behavior. What concerns the PSO dynamics, while the matter of consensus formation is well-studied, see, e.g., [CK02; OM98; YY15], only few theoretical statements regarding the properties of the found consensus are available. Besides the stochasticity and the usage of memory mechanisms of the method, the phenomenon of premature convergence observed for the basic PSO algorithm [van07; vdBE10] leads to a large number of variations and hard-to-analyze features of the method, thereby complicating the derivation of global convergence guarantees. For instance, a modified PSO version, so-called guaranteed convergence PSO was proposed in [vdBE10], which, however, also

²⁶In this section, we follow [PSO, Section 1].

only allows to prove convergence to local optima. In order to obtain therefrom a global search algorithm, the authors suggested adding purely stochastic particles to the swarm, which trivially makes the method capable of detecting a global minimizer but entails a computational time that coincides with the time required to examine every location in the search space, yielding an infeasible optimizer. Other works consider certain notions of weak convergence [BMW18] or provide probabilistic guarantees of finding locally optimal solutions [SW15]. Yet, a complete global numerical analysis of PSO was still lacking until our work [PSO]. For further references see, e.g., [Wit11; PSL11; ZWJ15] and the papers cited therein.

By casting the classical formulation of the PSO dynamics into a form that resembles the dynamics of CBO, see the description in the last paragraph of Section 2.1 for more details, the authors of [GP21] have paved the way for a rigorous mathematical convergence analysis of PSO. As we elaborated on in detail in Chapter 3, our approach leverages the analytical framework centered around the mean-field perspective. More precisely, adapting and transferring the technical analysis of [Car+18; Car+21] to the setting of PSO and thereby deriving convergence guarantees for PSO to global minimizers that are new to the literature of PSO, constitutes the central advancements made in [PSO]. Before presenting these results in the subsequent section, let us provide and explain the formulation of the PSO dynamics used for our analysis.

PSO with memory effects. Each individual particle of the swarm is described by a triplet (X^i, Y^i, V^i) ,²⁷ consisting of the position X^i , the historical best position Y^i , as well as the velocity V^i of the respective particle. While the memory mechanisms captured by the state variable Y^i are analogous to the ones of CBO with memory effects as thematized in Section 3.3, the second-order nature of the dynamics, i.e., the presence of the velocity V^i , is unique and characteristic of PSO. Let us now make the description of PSO as formulated in [GP21] more rigorous.

Given a finite time horizon $T > 0$ and a time discretization $0 < \Delta t < \dots < K\Delta t = T$ of $[0, T]$ with a suitable discrete time step size $\Delta t > 0$, we denote the position and velocity of the i th agent at time step k by $X_k^i \in \mathbb{R}^d$ and $V_k^i \in \mathbb{R}^d$, respectively, its historical best position stored in the particle's memory as $Y_k^i \in \mathbb{R}^d$, and the empirical measure of all agents' historical best positions at time step k by $\hat{\rho}_{Y,k}^N := \frac{1}{N} \sum_{i=1}^N \delta_{Y_k^i}$. For user-specified parameters $\alpha, \beta, \theta, \kappa, \gamma, m, \lambda_1, \sigma_1 > 0$ and $\lambda_2, \sigma_2 \geq 0$, the time-discrete evolution of the i th particle in the formulation of PSO with memory effects is given by the iterative update rule

$$X_k^i = X_{k-1}^i + \Delta t V_{k-1}^i, \quad (5.1a)$$

$$Y_k^i = Y_{k-1}^i + \Delta t \kappa \left(X_k^i - Y_{k-1}^i \right) S^{\beta, \theta} \left(X_k^i, Y_{k-1}^i \right), \quad (5.1b)$$

$$m V_k^i = m V_{k-1}^i - \Delta t \gamma V_{k-1}^i + \Delta t \lambda_1 \left(Y_k^i - X_k^i \right) + \Delta t \lambda_2 \left(y_\alpha^\mathcal{E}(\hat{\rho}_{Y,k}^N) - X_k^i \right) + \sigma_1 D \left(Y_k^i - X_k^i \right) B_k^{1,i} + \sigma_2 D \left(y_\alpha^\mathcal{E}(\hat{\rho}_{Y,k}^N) - X_k^i \right) B_k^{2,i}, \quad (5.1c)$$

²⁷Throughout this chapter, but limited to it, we denote by $((X_k^i)_{k=1, \dots, K})_{i=1, \dots, N}$ etc. the quantities of PSO (5.1) instead of standard CBO (2.2) or some of its variants.

where $((B_k^{m,i})_{k=1,\dots,K})_{i=1,\dots,N}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}$ for $m \in \{1, 2\}$. The system is complemented with independent initial data $(X_0^i, V_0^i, Y_0^i)_{i=1,\dots,N}$, typically such that $X_0^i = Y_0^i$ for all $i = 1, \dots, N$. The operator $S^{\beta,\theta}$ specifying the implementation of the historical best position Y^i is defined as in (3.118) in Section 3.3.

While the latter four terms in the update rule for the velocity V^i in (5.1c) are familiar to the reader since Section 3.3 and correspond to acceleration in the direction of the personal historical best of each particle as well as acceleration in the direction of the global consensus point $y_\alpha^\mathcal{E}$ computed on basis of the historical best positions of the particles, the first update in (5.1c) models friction with a coefficient commonly chosen as $\gamma = 1 - m \geq 0$, where $m > 0$ denotes the inertia weight.

The associated continuous-time mean-field dynamics of (5.1) is captured by the deterministic agent distribution $f \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d))$, which weakly satisfies the nonlinear nonlocal Vlasov-Fokker-Planck equation

$$\begin{aligned} \partial_t f_t + v \cdot \nabla_x f_t + \nabla_y \cdot \left(\kappa(x - y) S^{\beta,\theta}(x, y) f_t \right) \\ = \nabla_v \cdot \left(\frac{\gamma}{m} v f_t + \frac{\lambda_1}{m} (x - y) f_t + \frac{\lambda_2}{m} \left(x - y_\alpha^\mathcal{E}(\rho_{Y,t}) \right) f_t \right. \\ \left. + \left(\frac{\sigma_1^2}{2m^2} (D(x - y))^2 + \frac{\sigma_2^2}{2m^2} \left(D \left(x - y_\alpha^\mathcal{E}(\rho_{Y,t}) \right) \right)^2 \right) \nabla_v f_t \right), \end{aligned} \quad (5.2)$$

where the marginal law $\rho_{Y,t}$ is given by $\rho_Y(t, \bullet) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} df(t, x, \bullet, v)$.

PSO without memory effects. A reduced version of the PSO dynamics (5.1), which does not involve memory mechanisms and where each individual particle of the swarm is consequently described only by a tuple (X^i, V^i) , is given by²⁸

$$X_k^i = X_{k-1}^i + \Delta t V_{k-1}^i, \quad (5.3a)$$

$$m V_k^i = m V_{k-1}^i - \Delta t \gamma V_{k-1}^i + \Delta t \lambda \left(y_\alpha^\mathcal{E}(\widehat{\rho}_{X,k}^N) - X_k^i \right) + \sigma D \left(y_\alpha^\mathcal{E}(\widehat{\rho}_{X,k}^N) - X_k^i \right) B_k^{2,i}, \quad (5.3b)$$

where $\widehat{\rho}_{X,k}^N$ denotes the empirical measure of all agents' positions at time step k , i.e., $\widehat{\rho}_{X,k}^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}$. The associated continuous-time macroscopic description of (5.3) is captured by the deterministic agent distribution $f \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$, which weakly satisfies the nonlinear nonlocal Vlasov-Fokker-Planck equation

$$\partial_t f_t + v \cdot \nabla_x f_t = \nabla_v \cdot \left(\frac{\gamma}{m} v f_t + \frac{\lambda}{m} \left(x - y_\alpha^\mathcal{E}(\rho_{X,t}) \right) f_t + \frac{\sigma^2}{2m^2} \left(D \left(x - y_\alpha^\mathcal{E}(\rho_{X,t}) \right) \right)^2 \nabla_v f_t \right), \quad (5.4)$$

where the marginal law $\rho_{X,t}$ is given by $\rho_X(t, \bullet) = \int_{\mathbb{R}^d} df(t, \bullet, v)$.

²⁸Notice here, that we slightly abuse notations by not distinguishing in the notation between PSO with and without memory effects. However, it will be clear from the context, to what we refer.

5.2. Convergence of Particle Swarm Optimization to Global Minimizers²⁹

The theoretical contributions of [PSO] are concerned with the convergence of PSO with and without memory effects to global minimizers of the objective function \mathcal{E} , as we survey in what follows. Since we follow the analytical framework put forward by the authors of [Car+18; Car+21], which we addressed in the first paragraph of Section 3.1.1, we require a different set of assumptions.

Namely, we consider functions that satisfy in addition to Assumption 3.3 and A1 from Assumption 3.5 the following.

Assumption 5.1. In this section, we additionally require that $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ and that

- B1 $\|\nabla^2 \mathcal{E}\|_\infty \leq C_{\nabla^2 \mathcal{E}}$ for some constant $C_{\nabla^2 \mathcal{E}} > 0$, where $\|\cdot\|_\infty$ denotes the L^∞ -norm on $\mathcal{C}(\mathbb{R}^d)$,
- B2 there exist $\eta > 0$ and $\nu \in (0, \infty)$ such that for any $x \in \mathbb{R}^d$ there exists a global minimizer x^* of \mathcal{E} (which may depend on x) such that

$$\|x - x^*\|_2 \leq \frac{1}{\eta} (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu. \quad (5.5)$$

B1 is an additional regularity assumption about the function \mathcal{E} , which requires in particular that the objective is twice continuously differentiable. B2, on the other hand, is an inverse continuity property of the type A2, however, with a crucial difference in the details, which distinguishes the two main analytical frameworks described in Section 3.1.1 from one another. Unlike A2, B2 does neither require nor imply the uniqueness of the global minimizer x^* , but explicitly allows objective functions which have multiple global minimizers of identical quality, i.e., objective value $\underline{\mathcal{E}}$. For this reason, the convergence guarantees of this section are about convergence to global minimizers rather than global convergence to the global minimizer as are the ones in Chapter 3. In particular, the results presented in what follows must require restrictions about the initialization. An analysis in the framework of [CBO-I; CBO-II; CBO-IV] yielding global convergence guarantees for the class of objective functions familiar from Assumptions 3.3 and 3.5 is subject of ongoing work.

Under the aforementioned assumptions, we have the following statements about convergence of PSO with and without memory effects to global minimizers.

5.2.1. Convergence of Particle Swarm Optimization with Memory Effects to Global Minimizers³⁰

In the case of the PSO dynamics with memory effects, we obtain convergence of the with (5.1) associated continuous-time mean-field dynamics as follows.

²⁹In this section, we follow [PSO, Sections 2 to 4].

³⁰In this section, we follow [PSO, Section 3].

Theorem 5.2 (PSO with memory effects converges to global minimizers in the mean-field sense, [PSO, Theorem 4]). Let $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ satisfy W1–W2, A1, and B1. Let $(\bar{X}_t, \bar{Y}_t, \bar{V}_t)_{t \geq 0}$ denote a solution to the with the continuous-time variant of (5.1) associated self-consistent mean-field SDE according to Definition 3.2 (see also [PSO, Equation (1.8)]). Moreover, let us assume the well-preparation of the parameters together with the initial datum \bar{X}_0, \bar{Y}_0 and \bar{V}_0 in the sense that

P1 $\mu_1 > 0$ with

$$\mu_1 := \frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2} - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{4m^2} \right) \frac{12e^{-\alpha\underline{\mathcal{E}}}}{\mathbb{E} \exp(-\alpha\mathcal{E}(\bar{Y}_0))}, \quad (5.6)$$

P2 $\mu_2 > 0$ with

$$\begin{aligned} \mu_2 := & \frac{(\lambda_1 + \lambda_2)\gamma}{m^2} + \kappa\theta \left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2} \right) - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} - \frac{3\sigma_1^2}{2m^2} \\ & - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} \right) - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2} \right) \frac{24e^{-\alpha\underline{\mathcal{E}}}}{\mathbb{E} \exp(-\alpha\mathcal{E}(\bar{Y}_0))}, \end{aligned} \quad (5.7)$$

P3 it holds

$$\begin{aligned} & \left(\frac{\alpha\kappa m}{\lambda_1\chi} (C_{\nabla^2\mathcal{E}} + 2\alpha^2) + \frac{24C_{\nabla^2\mathcal{E}}^2\kappa}{\alpha\chi^3} \right) \frac{\mathbb{E}[\mathcal{H}(0)]}{\mathbb{E} \exp(-\alpha(\mathcal{E}(\bar{Y}_0) - \underline{\mathcal{E}}))} \\ & + \frac{6\kappa}{\alpha\chi} \frac{\mathbb{E} \|\nabla\mathcal{E}(\bar{X}_0)\|_2^2}{\mathbb{E} \exp(-\alpha(\mathcal{E}(\bar{Y}_0) - \underline{\mathcal{E}}))} < \frac{3}{32}, \end{aligned} \quad (5.8)$$

where

$$\chi := \frac{2}{5} \frac{\min\{\gamma/(2m), \mu_1, \mu_2\}}{\left((\gamma/(2m))^2 + 1 + 3\lambda_1/m + 2(\gamma/m)^2 \right)}. \quad (5.9)$$

Define the random variable

$$\begin{aligned} \mathcal{H}(t) := & \left(\frac{\gamma}{2m} \right)^2 \|\bar{X}_t - \mathbb{E}\bar{X}_t\|_2^2 + \frac{3}{2} \|\bar{V}_t\|_2^2 + \frac{1}{2} \left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2} \right) \|\bar{X}_t - \bar{Y}_t\|_2^2 \\ & + \frac{\gamma}{2m} \langle \bar{X}_t - \mathbb{E}\bar{X}_t, \bar{V}_t \rangle + \frac{\gamma}{m} \langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle. \end{aligned} \quad (5.10)$$

Then $\mathbb{E}[\mathcal{H}(t)]$ converges exponentially fast with rate χ to 0 as $t \rightarrow \infty$. Moreover, there exists some \tilde{x} , which may depend on α and f_0 , such that $\mathbb{E}\bar{X}_t \rightarrow \tilde{x}$ and $y_\alpha^\mathcal{E}(\rho_{Y,t}) \rightarrow \tilde{x}$ exponentially fast with rate $\chi/2$ as $t \rightarrow \infty$. Eventually, for any given accuracy $\varepsilon > 0$, there exists $\alpha_0 > 0$, which may depend on the dimension d , such that for all $\alpha > \alpha_0$, \tilde{x} satisfies $\mathcal{E}(\tilde{x}) - \underline{\mathcal{E}} \leq \varepsilon$. If \mathcal{E} additionally satisfies B2, we have $\|\tilde{x} - x^*\|_2 \leq \varepsilon^\nu/\eta$.

The proof of Theorem 5.2 is presented in [PSO, Section 3].

5.2.2. Convergence of Particle Swarm Optimization without Memory Effects to Global Minimizers³¹

In the case of the PSO dynamics without memory effects, we obtain an analogous statement for the with (5.3) associated continuous-time mean-field dynamics.

Theorem 5.3 (PSO without memory effects converges to global minimizers in the mean-field sense, [PSO, Theorem 2]). Let $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ satisfy W1–W2, A1, and B1. Let $(\bar{X}_t, \bar{V}_t)_{t \geq 0}$ denote a solution to the with the continuous-time variant of (5.3) associated self-consistent mean-field SDE according to Definition 3.2 (see also [PSO, Equation (2.3)]). Moreover, let us assume the well-preparation of the parameters together with the initial datum \bar{X}_0 and \bar{V}_0 in the sense that

P1 $\mu > 0$ with

$$\mu := \frac{\lambda\gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2} \right) \frac{4e^{-\alpha\underline{\mathcal{E}}}}{\mathbb{E} \exp(-\alpha\mathcal{E}(\bar{X}_0))}, \quad (5.11)$$

P2 it holds

$$\begin{aligned} & \frac{m\alpha}{2\gamma} \frac{\left(\mathbb{E} \left\langle \exp(-\alpha\mathcal{E}(\bar{X}_0)) \nabla \mathcal{E}(\bar{X}_0), \bar{V}_0 \right\rangle \right)_+}{\mathbb{E} \exp(-\alpha\mathcal{E}(\bar{X}_0))} \\ & + \frac{\alpha C_{\nabla^2 \mathcal{E}}}{\chi(\gamma/m - \chi)} \left(1 + \frac{8m\lambda}{\gamma^2} \right) \frac{\mathbb{E}[\mathcal{H}(0)]}{\left(\mathbb{E} \exp(-\alpha(\mathcal{E}(\bar{X}_0) - \underline{\mathcal{E}})) \right)^2} < \frac{3}{16}, \end{aligned} \quad (5.12)$$

with $x_+ = \max\{x, 0\}$ for $x \in \mathbb{R}$ denoting the positive part and where

$$\chi := \frac{2}{3} \frac{\min\{\gamma/m, \mu\}}{(\gamma/(2m))^2 + 1}. \quad (5.13)$$

Define the random variable

$$\mathcal{H}(t) := \left(\frac{\gamma}{2m} \right)^2 \left\| \bar{X}_t - \mathbb{E} \bar{X}_t \right\|^2 + \left\| \bar{V}_t \right\|^2 + \frac{\gamma}{2m} \left\langle \bar{X}_t - \mathbb{E} \bar{X}_t, \bar{V}_t \right\rangle. \quad (5.14)$$

Then $\mathbb{E}[\mathcal{H}(t)]$ converges exponentially fast with rate χ to 0 as $t \rightarrow \infty$. Moreover, there exists some \tilde{x} , which may depend on α and f_0 , such that $\mathbb{E} \bar{X}_t \rightarrow \tilde{x}$ and $x_\alpha^\mathcal{E}(\rho_{X,t}) \rightarrow \tilde{x}$ exponentially fast with rate $\chi/2$ as $t \rightarrow \infty$. Eventually, for any given accuracy $\varepsilon > 0$, there exists $\alpha_0 > 0$, which may depend on the dimension d , such that for all $\alpha > \alpha_0$, \tilde{x} satisfies $\mathcal{E}(\tilde{x}) - \underline{\mathcal{E}} \leq \varepsilon$. If \mathcal{E} additionally satisfies B2, we have $\|\tilde{x} - x^*\|_2 \leq \varepsilon^\nu/\eta$.

The proof of Theorem 5.3 is presented in [PSO, Section 2].

Without the presence of memory mechanisms, we can derive for the continuous-time dynamics of (5.3) (see also [PSO, Equation (2.1)]) a quantitative result about the mean-field approximation for PSO in the style of Section 3.1.2. This enables us to obtain

³¹In this section, we follow [PSO, Sections 2 and 4].

a holistic convergence statement for the numerical PSO method similar to the one of Section 3.1.3 for CBO. More precisely we have the following.

Theorem 5.4 (PSO without memory effects converges to global minimizers, [PSO, Theorem 6]). Let $\epsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Then, under the assumptions of Theorem 5.3, and with $K := T/\Delta t$, where $T = \mathcal{O}(\log(\tilde{\epsilon}^{-1})/\chi)$ with $\tilde{\epsilon}$ bounding the approximation error $\|\mathbb{E}\bar{X}_T - \tilde{x}\|$ thanks to Theorem 5.3 and for suitable $\Delta t > 0$, it holds for the discretized PSO dynamics (5.3) that

$$\left\| \frac{1}{N} \sum_{i=1}^N X_K^i - x^* \right\|_2^2 \leq \epsilon_{\text{total}} \quad (5.15)$$

with probability larger than

$$1 - \left(\delta + \epsilon_{\text{total}}^{-1} (C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \tilde{\epsilon} + \epsilon^{2\nu}/\eta^2) \right). \quad (5.16)$$

Here, m denotes the order of accuracy of the used discretization scheme (for the Euler-Maruyama scheme $m = 1/2$). Moreover, besides problem-dependent factors and the parameters of the method, the dependence of the constants is as follows. C_{NA} depends linearly on d and N , and exponentially on T . C_{MFA} depends exponentially on α , T and δ^{-1} . C_{LLN} depends on the moment bound from [PSO, Theorem 1]. Lastly, ϵ is chosen according to Theorem 5.3.

The proof of Theorem 5.4 is presented in [PSO, Section 4].

Chapter 6

Consensus-Based Optimization for Saddle Point Problems

In this final chapter before the conclusions, we showcase the last core contribution of this dissertation, for which we go beyond the task of solving optimization problems and, instead, turn towards tackling saddle point problems, i.e., finding global Nash equilibria. To this end, we present in [Section 6.1](#) a novel multi-particle metaheuristic derivative-free algorithm, consensus-based optimization for saddle point problems (CBO-SP), which we proposed in [\[CBO-SP\]](#) and which takes inspiration from the CBO method for optimization. It employs two groups of interacting particles, one of which performs a minimization over one variable while the other performs a maximization over the other variable. The two groups constantly exchange information through a suitably weighted average. This paradigm permits a passage to the mean-field limit and, as we sketch in [Section 6.2](#), makes the method amenable to theoretical analysis by allowing to obtain convergence guarantees under reasonable assumptions about the initialization and the objective function.

6.1. The Dynamics of Consensus-Based Optimization for Saddle Point Problems³²

Optimization problems where the goal is to find the best possible objective value for the worst-case scenario, so-called saddle point or minimax optimization problems are of the form

$$\min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x, y) \quad (6.1)$$

and attract a large amount of attention across several fields in applied mathematics and beyond. This includes classical applications in game theory [\[vNM07\]](#), economics [\[Mye91\]](#), engineering and signal processing [\[Goh+09; LDL13; Cha+20\]](#), but also cutting-edge topics in machine learning such as the training of GANs [\[Goo+20\]](#), adversarial training [\[Mad+18b\]](#), and fair machine learning [\[Mad+18a\]](#). In most of the applications of recent interest, the payoff function \mathcal{E} is nonconvex-nonconcave, making the problem of finding a global equilibrium in the sense of [Definition 6.1](#) in general NP-hard [\[MK87\]](#) and the available toolset and theories very limited, see, e.g., the review paper [\[Raz+20\]](#).

³²In this section, we follow [\[CBO-SP, Section 1\]](#).

A well-known notion of optimality originating from game theory is the one of Nash equilibria (also referred to as saddle points) [Nas50], where neither of the players has anything to gain by changing only its own strategy.

Definition 6.1. A point $(x^*, y^*) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ is called Nash equilibrium or saddle point of a function \mathcal{E} if it holds

$$\mathcal{E}(x^*, y) \leq \mathcal{E}(x^*, y^*) \leq \mathcal{E}(x, y^*) \quad \text{for all } (x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \quad (6.2)$$

or, equivalently, if

$$\min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x, y) = \mathcal{E}(x^*, y^*) = \max_{y \in \mathbb{R}^{d_2}} \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y). \quad (6.3)$$

To keep the notation concise we write \mathcal{E}^* for $\mathcal{E}(x^*, y^*)$ in what follows.

Constituting the core contribution of our work [CBO-SP], we propose a novel zero-order consensus-based optimization method for finding the global Nash equilibrium (x^*, y^*) of a smooth objective function $\mathcal{E} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$, which is designed to be amenable to a rigorous theoretical convergence analysis, missing so far in the literature on population-based methods for minimax problems [SK02; KHdS04; LPV02]. In contrast to gradient descent-ascent-like algorithms such as [Bub15; Haz16; Nou+19; Raz+20; Com+24], CBO-SP is derivative-free. Taking inspiration from CBO for optimization, it employs two sets of particles X^1, \dots, X^{N_1} and Y^1, \dots, Y^{N_2} , one for minimization, the other for maximization, with the aim of exploring the domain and forming a consensus about the location of the saddle point (x^*, y^*) . More formally, the strategy is as follows.

Given a finite time horizon $T > 0$ and a time discretization $0 < \Delta t < \dots < K\Delta t = T$ of $[0, T]$ with a suitable discrete time step size $\Delta t > 0$, we denote the position of the respective i th agent at time step k by $X_k^i \in \mathbb{R}^{d_1}$ and $Y_k^i \in \mathbb{R}^{d_2}$, and their associated empirical measures by $\hat{\rho}_{X,k}^{N_1}$ and $\hat{\rho}_{Y,k}^{N_2}$. For user-specified parameters $\alpha, \beta, \lambda_1, \sigma_1, \lambda_2, \sigma_2 > 0$, the time-discrete evolution of CBO-SP is given by the iterative update rule

$$X_k^i = X_{k-1}^i - \lambda_1 \Delta t \left(X_{k-1}^i - x_\alpha^{\mathcal{E}, Y}(\hat{\rho}_{X,k-1}^{N_1}) \right) + \sigma_1 D \left(X_{k-1}^i - x_\alpha^{\mathcal{E}, Y}(\hat{\rho}_{X,k-1}^{N_1}) \right) B_k^{X,i}, \quad (6.4a)$$

$$Y_k^i = Y_{k-1}^i - \lambda_2 \Delta t \left(Y_{k-1}^i - y_\beta^{\mathcal{E}, X}(\hat{\rho}_{Y,k-1}^{N_2}) \right) + \sigma_2 D \left(Y_{k-1}^i - y_\beta^{\mathcal{E}, X}(\hat{\rho}_{Y,k-1}^{N_2}) \right) B_k^{Y,i}, \quad (6.4b)$$

where $((B_k^{X,i})_{k=1, \dots, K})_{i=1, \dots, N_1}$ and $((B_k^{Y,i})_{k=1, \dots, K})_{i=1, \dots, N_2}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^{d_1} and \mathbb{R}^{d_2} , respectively, with zero mean and covariance matrix $\Delta t \text{Id}$. The consensus point is computed as

$$x_\alpha^{\mathcal{E}, Y}(\hat{\rho}_{X,k}^{N_1}) = \int x \frac{\omega_\alpha^\mathcal{E}(x, \int y d\hat{\rho}_{Y,k}^{N_2}(y))}{\|\omega_\alpha^\mathcal{E}(\bullet, \int y d\hat{\rho}_{Y,k}^{N_2}(y))\|_{L_1(\hat{\rho}_{X,k}^{N_1})}} d\hat{\rho}_{X,k}^{N_1}(x), \quad (6.5a)$$

$$y_\beta^{\mathcal{E}, X}(\hat{\rho}_{Y,k}^{N_2}) = \int y \frac{\omega_{-\beta}^\mathcal{E}(\int x d\hat{\rho}_{X,k+1}^{N_1}(x), y)}{\|\omega_{-\beta}^\mathcal{E}(\int x d\hat{\rho}_{X,k+1}^{N_1}(x), \bullet)\|_{L_1(\hat{\rho}_{Y,k}^{N_2})}} d\hat{\rho}_{Y,k}^{N_2}(y). \quad (6.5b)$$

The associated continuous-time mean-field dynamics of (6.4), which is the basis for the analytical considerations discussed in the subsequent section, is captured by the deterministic agent distribution $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^{d_1+d_2}))$, which weakly satisfies the nonlinear nonlocal Fokker-Planck equation

$$\begin{aligned} \partial_t \rho_t &= \lambda_1 \operatorname{div}_x \left(\left(x - x_\alpha^{\mathcal{E}, Y}(\rho_t^X) \right) \rho_t \right) + \lambda_2 \operatorname{div}_y \left(\left(y - y_\beta^{\mathcal{E}, X}(\rho_t^Y) \right) \rho_t \right) \\ &+ \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \partial_{x_k x_k}^2 \left(\left(x - x_\alpha^{\mathcal{E}, Y}(\rho_t^X) \right)_k^2 \rho_t \right) + \frac{\sigma_2^2}{2} \sum_{k=1}^{d_2} \partial_{y_k y_k}^2 \left(\left(y - y_\beta^{\mathcal{E}, X}(\rho_t^Y) \right)_k^2 \rho_t \right) \end{aligned} \quad (6.6)$$

with marginal laws $\rho_{X,t}$ and $\rho_{Y,t}$ given by $\rho_X(t, \bullet) = \int_{\mathbb{R}^{d_2}} d\rho(t, \bullet, y)$ and $\rho_Y(t, \bullet) = \int_{\mathbb{R}^{d_1}} d\rho(t, x, \bullet)$, respectively.

6.2. Convergence of Consensus-Based Optimization for Saddle Point Problems to Saddle Points³³

Besides proposing the CBO-SP algorithm (6.4) in [CBO-SP], we provide a mathematical perspective on its behavior by taking the familiar mean-field point of view, i.e., investigating the associated mean-field dynamics (6.6).

For this, we consider functions that satisfy the following.

Assumption 6.2. Throughout this section we are interested in objective functions $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^{d_1+d_2})$, for which

S1 there exist two functions $\underline{\mathcal{E}} \in \mathcal{C}^1(\mathbb{R}^{d_2})$ and $\bar{\mathcal{E}} \in \mathcal{C}^1(\mathbb{R}^{d_1})$ such that

$$\underline{\mathcal{E}}(y) \leq \mathcal{E}(x, y) \leq \bar{\mathcal{E}}(x) \quad \text{for all } (x, y) \in \mathbb{R}^{d_1+d_2}. \quad (6.7)$$

The functions $\underline{\mathcal{E}}$ and $\bar{\mathcal{E}}$ shall, for a constant $\bar{C}_{\nabla \mathcal{E}} > 0$, satisfy $\|\nabla \underline{\mathcal{E}}(y)\|_2 \leq \bar{C}_{\nabla \mathcal{E}}$ for all $y \in \mathbb{R}^{d_2}$ and $\|\nabla \bar{\mathcal{E}}(x)\|_2 \leq \bar{C}_{\nabla \mathcal{E}}$ for all $x \in \mathbb{R}^{d_1}$.

S2 there exists a constant $C_1 > 0$ such that it holds

$$\begin{aligned} |\mathcal{E}(x, y) - \mathcal{E}(x', y')| &\leq C_1 \left(1 + \|x\|_2 + \|x'\|_2 + \|y\|_2 + \|y'\|_2 \right) \\ &\cdot (\|x - x'\|_2 + \|y - y'\|_2) \end{aligned} \quad (6.8)$$

for all $(x, y), (x', y') \in \mathbb{R}^{d_1+d_2}$ and $s \in [0, 1]$.

S3 there exists a constant $C_2 > 0$ obeying

$$\mathcal{E}(x, y) - \underline{\mathcal{E}}(y + sy') \leq C_2(1 + \|x\|_2^2 + \|y\|_2^2 + \|y'\|_2^2), \quad (6.9a)$$

and

$$\bar{\mathcal{E}}(x + sx') - \mathcal{E}(x, y) \leq C_2(1 + \|x\|_2^2 + \|x'\|_2^2 + \|y\|_2^2) \quad (6.9b)$$

for all $(x, y), (x', y') \in \mathbb{R}^{d_1+d_2}$ and $s \in [0, 1]$.

³³In this section, we follow [CBO-SP, Section 3].

S4 there exists a constant $C_{\nabla\mathcal{E}} > 0$ such that

$$\max \left\{ \sup_{(x,y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} \|\nabla_x \mathcal{E}(x,y)\|_2, \sup_{(x,y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} \|\nabla_y \mathcal{E}(x,y)\|_2 \right\} \leq C_{\nabla\mathcal{E}}, \quad (6.10)$$

and a constant $C_{\nabla^2\mathcal{E}} > 0$ such that

$$\max \left\{ \max_{k=1,\dots,d_1} \|\partial_{x_k x_k}^2 \mathcal{E}\|_\infty, \max_{k=1,\dots,d_2} \|\partial_{y_k y_k}^2 \mathcal{E}\|_\infty, \|\rho(\nabla_x^2 \mathcal{E})\|_\infty, \|\rho(\nabla_y^2 \mathcal{E})\|_\infty \right\} \leq C_{\nabla^2\mathcal{E}}, \quad (6.11)$$

where $\|\cdot\|_\infty$ denotes the L^∞ -norm on $\mathcal{C}(\mathbb{R}^{d_1+d_2})$ and ρ denotes the spectral radius.

S5 there exist constants $\epsilon_0, \eta, \nu > 0$ such that for each $(x,y) \in \mathbb{R}^{d_1+d_2}$ satisfying $\mathcal{E}^* - \mathcal{E}(x^*, y) \leq \epsilon_0$ and $\mathcal{E}(x, y^*) - \mathcal{E}^* \leq \epsilon_0$ for some saddle point (x^*, y^*) of \mathcal{E} , we have

$$\|x - x^*\|_2 \leq \frac{1}{\eta} (|\mathcal{E}(x, y^*) - \mathcal{E}^*|)^\nu \quad \text{and} \quad \|y - y^*\|_2 \leq \frac{1}{\eta} (|\mathcal{E}(x^*, y) - \mathcal{E}^*|)^\nu. \quad (6.12)$$

S1–S3 are boundedness and growth conditions on \mathcal{E} , which are in particular sufficient to ensure the well-posedness of the continuous-time version of (6.4) as well as its associated mean-field dynamics (6.6), see, e.g., [CBO-SP, Section 2]. S4 comprises mere technical regularity assumptions in terms of the first and second derivatives. In particular, it requires that the gradients as well as second-order derivatives of \mathcal{E} are uniformly bounded, which is, however, necessary only for theoretical analysis of the long-term behavior of the algorithm. As a purely zero-order derivative-free method, our CBO-SP algorithm requires in practice only point-wise evaluations of \mathcal{E} . S5, on the other hand, should be regarded as a tractability condition on the landscape of the objective function \mathcal{E} . It imposes coercivity of \mathcal{E} around saddle points, which relates the distance from (x^*, y^*) with the value of the objective function.

Under these assumptions, we have the following consensus and convergence statements for CBO-SP provided certain well-preparedness conditions about the initialization and the parameters of the method are met as indicated in the statement.

Theorem 6.3 (CBO-SP converges to saddle points in the mean-field sense, [CBO-SP, Theorem 11]). Let $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^{d_1+d_2})$ satisfy S1–S4. Let $(\bar{X}_t, \bar{Y}_t)_{t \geq 0}$ denote a solution to the with the continuous-time variant of (6.4) associated self-consistent mean-field SDE according to Definition 3.2 (see also [CBO-SP, Equation (3)]). Define the functionals

$$\text{Var}^X(t) = \mathbb{E} \left\| \bar{X}_t - \mathbb{E} \bar{X}_t \right\|_2^2 \quad \text{and} \quad \text{Var}^Y(t) = \mathbb{E} \left\| \bar{Y}_t - \mathbb{E} \bar{Y}_t \right\|_2^2. \quad (6.13)$$

Then the following statements hold.

- (1) Under the assumption of well-preparedness of the initial datum (\bar{X}_0, \bar{Y}_0) and the parameters $\alpha, \beta, \lambda_1, \lambda_2, \sigma_1$ and σ_2 in the sense of [CBO-SP, Definition 10, P1–P3], Var^X and Var^Y converge exponentially fast to 0 as $t \rightarrow \infty$. More precisely, for rates $\mu_1, \mu_2 > 0$ defined as in [CBO-SP, Definition 10, P1], it holds

$$\text{Var}^X(t) + \text{Var}^Y(t) \leq \text{Var}^X(0)e^{-\mu_1 t} + \text{Var}^Y(0)e^{-\mu_2 t}. \quad (6.14)$$

Moreover, there exists some (\tilde{x}, \tilde{y}) depending in particular on α and β such that, as $t \rightarrow \infty$,

$$(\mathbb{E}\bar{X}_t, \mathbb{E}\bar{Y}_t) \rightarrow (\tilde{x}, \tilde{y}) \quad \text{and} \quad (x_\alpha^{\mathcal{E}, Y}(\rho_{X,t}), y_\beta^{\mathcal{E}, X}(\rho_{Y,t})) \rightarrow (\tilde{x}, \tilde{y}). \quad (6.15)$$

- (2) For any given accuracy $\varepsilon > 0$, there exist some $\alpha_0, \beta_0 > 0$ such that for all $\alpha \geq \alpha_0$ and $\beta \geq \beta_0$ the point (\tilde{x}, \tilde{y}) from (1) (which may depend on α and β) satisfies

$$|\mathcal{E}(\tilde{x}, \tilde{y}) - \mathcal{E}^*| \leq \varepsilon \quad \text{as well as} \quad \mathcal{E}^* - \mathcal{E}(x^*, \tilde{y}) \leq \varepsilon \quad \text{and} \quad \mathcal{E}(\tilde{x}, y^*) - \mathcal{E}^* \leq \varepsilon \quad (6.16)$$

provided that the well-preparedness assumptions [CBO-SP, Definition 10, P1–P4] hold for such α and β together with the initial datum (\bar{X}_0, \bar{Y}_0) .

- (3) If \mathcal{E} additionally satisfies S4 with respect to (\tilde{x}, \tilde{y}) from (2) with $\varepsilon \leq \varepsilon_0$, i.e., there exists some saddle point (x^*, y^*) depending on (\tilde{x}, \tilde{y}) such that $\|\tilde{x} - x^*\|_2 \leq (|\mathcal{E}(\tilde{x}, y^*) - \mathcal{E}^*|)^\nu / \eta$ and $\|\tilde{y} - y^*\|_2 \leq (|\mathcal{E}(x^*, \tilde{y}) - \mathcal{E}^*|)^\nu / \eta$, then we have

$$\|(\tilde{x}, \tilde{y}) - (x^*, y^*)\|_2 \leq \frac{2}{\eta} \varepsilon^\nu \quad (6.17)$$

provided that the well-preparedness assumptions [CBO-SP, Definition 10, P1–P4] hold for sufficiently large α and β together with the initial datum (\bar{X}_0, \bar{Y}_0) .

The proof of Theorem 6.3 is presented in [CBO-SP, Section 4].

Chapter 7

Conclusions

This dissertation laid mathematical foundations for the numerical analysis of interacting multi-particle systems in the setting of nonconvex nonsmooth optimization in high dimensions.

It is based on the publications listed at the end of the preface.

Interacting multi-particle methods are part of the vast class of heuristics and metaheuristics, which comprise evolution strategies, evolutionary programming methods, genetic algorithms, PSO, random search, the Nelder-Mead simplex heuristic, the Metropolis-Hastings algorithm, simulated annealing, and many more well-known and well-established methods. In contrast to the classical optimization paradigm, which relies on and mostly exploits local information about the objective function (examples being gradient descent-like, Newton-like, or trust region methods), many metaheuristics intertwine local improvement procedures and deterministic decisions with global strategies and stochastic processes, and employ a system of interacting particles to explore the parameter space and to consecutively exploit the gathered information through communication, with the overall goal to design an efficient yet effective procedure for reliably and robustly finding globally optimal solutions.

Despite their tremendous empirical success, broad spectrum of applicability, ease of handling, and wide use in practice, their rigorous theoretical analysis has largely remained elusive. This is mostly attributed to the inherent complexity and intricacies of the particle system arising from the nonlinear and nontrivial working principles and interaction rules underlying the algorithm, the prevalent stochasticity, and a possibly large number of involved agents, in particular in the case of hard nonconvex and high-dimensional problems.

However, given the necessity of capable, reliable, and robust algorithms that come with informative and solid convergence guarantees, a mathematical analysis framework for these methods that allows to derive rigorous quantitative estimates about the finite-time behavior of the algorithms with explicit rates of convergence, is of crucial interest and indispensable to warrant their applicability in particular in security-, privacy-, and fairness-sensitive applications. To close this gap, we covered in this dissertation algorithms designed for classical global optimization problems as well as saddle point or so-called minimax optimization problems. Tasks of either type are of fundamental interest throughout science and engineering, including most recent developments in machine learning and artificial intelligence. We further established a surprising and interesting,

yet, so far, largely unexplored and unexploited link between the derivative-free and the gradient-based world in optimization.

Mathematical foundations of interacting multi-particle methods. Due to the aforementioned inherent complexity and intricacies of large systems of interacting particles, a mean-field perspective is taken at the heart of our analysis, which, as we elaborated on in the introduction, has become a prominent and fruitful theoretical avenue. In this vein, the central observations and core contributions of this dissertation are connected to and rest on insights gained in the infinite particle regime. With original complexities of the objective function being provably alleviated on this level and the hardness of the original optimization problem having disappeared, this point of view enabled us to understand, unveil, and distill those key internal mechanisms of the investigated purpose-driven interacting particle systems that are centrally and crucially responsible for the, in a wide variety of applications, empirically observed successes. However, in order to infer properties of practical interest about the associated implemented multi-agent algorithms, we went beyond the investigated mean-field descriptions by providing quantitative mean-field approximation results. This yielded holistic convergence proofs in form of probabilistic global convergence guarantees for the interacting particle systems under investigation. While we focused in this dissertation on the CBO and PSO algorithms as well as some of their variants, the general analytical framework is flexible and versatile enough to be adapted and extended to a wider class of numerical algorithms, as is demonstrated convincingly by the recent literature in this field.

A link between the derivative-free and gradient-based worlds in optimization. By viewing CBO from a different angle, we paved the way for a completely novel analytical approach to theoretically investigate gradient-based learning algorithms that are considered one of the cornerstones of the astounding successes of machine learning. Forging such an unexpected, yet, surprising and intriguing link between these two, up to now, rather separated worlds in optimization, will enable us to drive forward our theoretical understanding of both gradient-based learning methods and metaheuristic black-box optimization algorithms. We further widen the scope of applications of methods which—in one way or another, be it explicitly or implicitly—estimate and exploit gradients. In particular, we believe these insights to bear the potential for designing efficient and reliable training methods which behave like first-order methods while not relying on the ability of computing gradients.

A word on the future. Plenty of exciting and relevant research directions that revolve around the topic of interacting multi-particle systems for optimization are left for further investigations.

First of all, the general mathematical framework with a mean-field perspective at its heart, which we thematized in this dissertation, beckons to be applied to other classical, recent, and yet-to-be-designed heuristics and metaheuristics.

On this note, a taxonomy for metaheuristics, in general, requires a closer study to determine the relationships among them in a broader fashion.

Secondly, while PSO is well-established across several communities in academia and industry, CBO and its variants still have to prove themselves competitive in the relevant benchmark tests of interest in order to be appealing to these communities. For this, in particular, parallel implementations that further improve the computational complexity and efficiency are of crucial importance.

From a theoretical perspective, a better understanding of the hyperparameters of the method, in particular the parameter α and its relationship with the intrinsic hardness of the optimization problem, is of fundamental interest. This will give deeper insights into the question, for which classes of objective functions CBO, PSO, and related methods, as well as metaheuristics in general, are effective and efficient.

Moreover, this might inspire the design of diffusions that are less agnostic to and rather adaptive w.r.t. the objective function \mathcal{E} , leading to a more capable scheme in case of ill-conditioned problems.

Thirdly, suitable variations of CBO may be worth to be introduced for robust, bilevel, stochastic, and infinite-dimensional optimizations.

Bibliography

- [AK89] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines. A stochastic approach to combinatorial optimization and neural computing*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Ltd., Chichester, 1989, pp. xii+272.
- [AO11] D. Acemoglu and A. E. Ozdaglar. “Opinion Dynamics and Learning in Social Networks.” In: *Dyn. Games Appl.* 1.1 (2011), pp. 3–49.
- [Aga+11] A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin. “Stochastic convex optimization with bandit feedback.” In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Vol. 24. Curran Associates, Inc., 2011.
- [Aki+10] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi. “Bidirectional Relation between CMA Evolution Strategies and Natural Evolution Strategies.” In: *Parallel Problem Solving from Nature - PPSN XI, 11th International Conference, Kraków, Poland, September 11-15, 2010, Proceedings, Part I*. Ed. by R. Schaefer, C. Cotta, J. Kolodziej, and G. Rudolph. Vol. 6238. Lecture Notes in Computer Science. Springer, 2010, pp. 154–163.
- [Alb+16] G. Albi, M. Bongini, E. Cristiani, and D. Kalise. “Invisible control of self-organizing agents leaving unknown environments.” In: *SIAM J. Appl. Math.* 76.4 (2016), pp. 1683–1710.
- [AFT23] G. Albi, F. Ferrarese, and C. Totzeck. “Kinetic based optimization enhanced by genetic dynamics.” In: *arXiv preprint arXiv:2306.09199* (2023).
- [APU23] K. Althaus, I. Papaioannou, and E. Ullmann. “Consensus-based rare event estimation.” In: *arXiv preprint arXiv:2304.09077* (2023).
- [AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Second. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008, pp. x+334.
- [Ani00] M. Anitescu. “Degenerate nonlinear programming with a quadratic growth condition.” In: *SIAM J. Optim.* 10.4 (2000), pp. 1116–1135.
- [AOY19] D. Araújo, R. I. Oliveira, and D. Yukimura. “A mean-field limit for certain deep neural networks.” In: *arXiv preprint arXiv:1906.00193* (2019).
- [AGR00] H. Attouch, X. Goudou, and P. Redont. “The heavy ball with friction method. I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system.” In: *Commun. Contemp. Math.* 2.1 (2000), pp. 1–34.

- [BFM97] T. Bäck, D. B. Fogel, and Z. Michalewicz, eds. *Handbook of evolutionary computation*. Institute of Physics Publishing, Bristol; Oxford University Press, New York, 1997, pp. xviii+1113.
- [Bae+22] H.-O. Bae, S.-Y. Ha, M. Kang, H. Lim, C. Min, and J. Yoo. “A constrained consensus based optimization algorithm and its application to finance.” In: *Appl. Math. Comput.* 416 (2022), Paper No. 126726, 10.
- [CBX] R. Bailo, A. Barbaro, S. N. Gomes, K. Riedl, T. Roith, C. Totzeck, and U. Vaes. “CBX: Python and Julia packages for consensus-based interacting particle methods.” In: *arXiv preprint arXiv:2403.14470* (2024).
- [Bak96] P. Bak. *How nature works*. The science of self-organized criticality. Copernicus, New York, 1996, pp. xiv+212.
- [BBP22] A. Benfenati, G. Borghi, and L. Pareschi. “Binary interaction methods for high dimensional global optimization and machine learning.” In: *Appl. Math. Optim.* 86.1 (2022), Paper No. 9, 41.
- [BGL05] M. Benzi, G. H. Golub, and J. Liesen. “Numerical solution of saddle point problems.” In: *Acta Numer.* 14 (2005), pp. 1–137.
- [Ber+11] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. “Algorithms for Hyper-Parameter Optimization.” In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Vol. 24. Curran Associates, Inc., 2011.
- [Bia+09] L. Bianchi, M. Dorigo, L. M. Gambardella, and W. J. Gutjahr. “A survey on metaheuristics for stochastic combinatorial optimization.” In: *Nat. Comput.* 8.2 (2009), pp. 239–287.
- [BR03] C. Blum and A. Roli. “Metaheuristics in combinatorial optimization: Overview and conceptual comparison.” In: *ACM Comput. Surv.* 35.3 (2003), pp. 268–308.
- [Bol+17] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. “From error bounds to the complexity of first-order descent methods for convex functions.” In: *Math. Program.* 165.2, Ser. A (2017), pp. 471–507.
- [Bol77] L. Boltzmann. *Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht*. Kk Hof- und Staatsdruckerei, 1877.
- [Bor24] G. Borghi. “Mean-field theory for consensus-based optimization and extensions to constrained and multi-objective problems.” PhD thesis. RWTH Aachen University, 2024.
- [Bor23] G. Borghi. “Repulsion dynamics for uniform Pareto front approximation in multi-objective optimization problems.” In: *PAMM* 23.1 (2023), e202200285.

- [BGP23] G. Borghi, S. Grassi, and L. Pareschi. “Consensus based optimization with memory effects: random selection and applications.” In: *Chaos Solitons Fractals* 174 (2023), Paper No. 113859, 17.
- [BH23] G. Borghi and M. Herty. “Model predictive control strategies using consensus-based optimization.” In: *arXiv preprint arXiv:2312.13085* (2023).
- [BHP22] G. Borghi, M. Herty, and L. Pareschi. “A consensus-based algorithm for multi-objective optimization and its mean-field description.” In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE. 2022, pp. 4131–4136.
- [BHP23a] G. Borghi, M. Herty, and L. Pareschi. “An adaptive consensus based method for multi-objective optimization with uniform Pareto front approximation.” In: *Appl. Math. Optim.* 88.2 (2023), Paper No. 58, 43.
- [BHP23b] G. Borghi, M. Herty, and L. Pareschi. “Constrained consensus-based optimization.” In: *SIAM J. Optim.* 33.1 (2023), pp. 211–236.
- [BP23] G. Borghi and L. Pareschi. “Kinetic description and convergence analysis of genetic algorithms for global optimization.” In: *arXiv preprint arXiv:2310.08562* (2023).
- [BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004, pp. xiv+716.
- [BC61] D. Brouwer and G. M. Clemence. *Methods of celestial mechanics*. Academic Press, New York-London, 1961, pp. xii+598.
- [Bro67] C. G. Broyden. “Quasi-Newton methods and their application to function minimisation.” In: *Math. Comp.* 21 (1967), pp. 368–381.
- [BMW18] V. Bruned, A. Mas, and S. Wlodarczyk. “Weak convergence of particle swarm optimization.” In: *arXiv preprint arXiv:1811.04924* (2018).
- [Brü21] T. Brünnette. *Consensus Based Optimization with Finite Range Interaction*. 2021.
- [Bub15] S. Bubeck. “Convex Optimization: Algorithms and Complexity.” In: *Found. Trends Mach. Learn.* 8.3-4 (2015), pp. 231–357.
- [BWR22] L. Bungert, P. Wacker, and T. Roith. “Polarized consensus-based dynamics for optimization and sampling.” In: *arXiv preprint arXiv:2211.05238* (2022).
- [BHW24] J. Byeon, S.-Y. Ha, and J.-H. Won. “Discrete Consensus-Based Optimization.” In: *arXiv preprint arXiv:2403.03430* (2024).
- [CRT06] E. J. Candès, J. K. Romberg, and T. Tao. “Stable signal recovery from incomplete and inaccurate measurements.” In: *Comm. Pure Appl. Math.* 59.8 (2006), pp. 1207–1223.

- [Car+18] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. “An analytical framework for consensus-based global optimization method.” In: *Math. Models Methods Appl. Sci.* 28.6 (2018), pp. 1037–1066.
- [Car+22] J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes. “Consensus-based sampling.” In: *Stud. Appl. Math.* 148.3 (2022), pp. 1069–1140.
- [Car+21] J. A. Carrillo, S. Jin, L. Li, and Y. Zhu. “A consensus-based global optimization method for high dimensional machine learning problems.” In: *ESAIM Control Optim. Calc. Var.* 27.suppl. (2021), Paper No. S5, 22.
- [Car+23] J. A. Carrillo, N. G. Trillos, S. Li, and Y. Zhu. “FedCBO: Reaching Group Consensus in Clustered Federated Learning through Consensus-based Optimization.” In: *arXiv preprint arXiv:2305.02894* (2023).
- [CCH14] J. A. Carrillo, Y.-P. Choi, and M. Hauray. “The derivation of swarming models: mean-field limit and Wasserstein distances.” In: *Collective dynamics from bacteria to crowds*. Vol. 553. CISM Courses and Lect. Springer, Vienna, 2014, pp. 1–46.
- [CTV23] J. A. Carrillo, C. Totzeck, and U. Vaes. “Consensus-based optimization and ensemble Kalman inversion for global optimization problems with constraints.” In: *Modeling and Simulation for Collective Dynamics*. World Scientific, 2023, pp. 195–230.
- [CD22a] L.-P. Chaintron and A. Diez. “Propagation of chaos: a review of models, methods and applications. I. Models and methods.” In: *Kinet. Relat. Models* 15.6 (2022), pp. 895–1015.
- [CD22b] L.-P. Chaintron and A. Diez. “Propagation of chaos: a review of models, methods and applications. II. Applications.” In: *Kinet. Relat. Models* 15.6 (2022), pp. 1017–1173.
- [Cha+20] T. Chang, M. Hong, H. Wai, X. Zhang, and S. Lu. “Distributed Learning in the Nonconvex World: From batch data to streaming and beyond.” In: *IEEE Signal Process. Mag.* 37.3 (2020), pp. 26–38.
- [CJL22] J. Chen, S. Jin, and L. Lyu. “A consensus-based global optimization method with adaptive momentum estimation.” In: *Commun. Comput. Phys.* 31.4 (2022), pp. 1296–1316.
- [Che+17] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh. “ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models.” In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*. Ed. by B. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, and A. Sinha. ACM, 2017, pp. 15–26.
- [CHQ23] E. Chenchene, H. Huang, and J. Qiu. “A consensus-based algorithm for non-convex multiplayer games.” In: *arXiv preprint arXiv:2311.08270* (2023).

- [Che23] S. Chewi. “An optimization perspective on log-concave sampling and beyond.” PhD thesis. Massachusetts Institute of Technology, 2023.
- [Chi+23] P. Chiang, R. Ni, D. Y. Miller, A. Bansal, J. Geiping, M. Goldblum, and T. Goldstein. “Loss Landscapes are All You Need: Neural Network Generalization Can Be Explained Without the Implicit Bias of Gradient Descent.” In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [CHS87] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. “Diffusion for global optimization in \mathbb{R}^n .” In: *SIAM J. Control Optim.* 25.3 (1987), pp. 737–753.
- [Chi22] L. Chizat. “Mean-Field Langevin Dynamics: Exponential Convergence and Annealing.” In: *Trans. Mach. Learn. Res.* 2022 (2022).
- [CB18] L. Chizat and F. Bach. “On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.
- [Chi+22] L. Chizat, M. Colombo, X. Fernández-Real, and A. Figalli. “Infinite-width limit of deep linear neural networks.” In: *arXiv preprint arXiv:2211.16980* (2022).
- [CHQ22] C. Cipriani, H. Huang, and J. Qiu. “Zero-inertia limit: from particle swarm optimization to consensus-based optimization.” In: *SIAM J. Math. Anal.* 54.3 (2022), pp. 3091–3121.
- [CK02] M. Clerc and J. Kennedy. “The particle swarm - explosion, stability, and convergence in a multidimensional complex space.” In: *IEEE Trans. Evol. Comput.* 6.1 (2002), pp. 58–73.
- [CGP05] G. M. Coclite, M. Garavello, and B. Piccoli. “Traffic flow on a road network.” In: *SIAM J. Math. Anal.* 36.6 (2005), pp. 1862–1886.
- [Com+24] E. M. Compagnoni, A. Orvieto, H. Kersting, F. N. Proske, and A. Lucchi. “SDEs for Minimax Optimization.” In: *arXiv preprint arXiv:2402.12508* (2024).
- [CST97] A. R. Conn, K. Scheinberg, and P. L. Toint. “Recent progress in unconstrained nonlinear optimization without derivatives.” In: vol. 79. 1-3. Lectures on mathematical programming (ismp97) (Lausanne, 1997). 1997, pp. 397–414.
- [CGT00] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-region methods*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2000, pp. xx+959.

- [Cou+05] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. “Effective leadership and decision-making in animal groups on the move.” In: *Nature* 433.7025 (2005), pp. 513–516.
- [CPT11] E. Cristiani, B. Piccoli, and A. Tosin. “Multiscale modeling of granular flows with application to crowd dynamics.” In: *Multiscale Model. Simul.* 9.1 (2011), pp. 155–182.
- [De 93] E. De Giorgi. “New problems on minimizing movements.” In: *Boundary value problems for partial differential equations and applications*. Vol. 29. RMA Res. Notes Appl. Math. Masson, Paris, 1993, pp. 81–98.
- [Dea+12] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. Le, and A. Ng. “Large Scale Distributed Deep Networks.” In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger. Vol. 25. Curran Associates, Inc., 2012.
- [Def+00] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. “Mixing beliefs among interacting agents.” In: *Adv. Complex Syst.* 3.1-4 (2000), pp. 87–98.
- [DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Second. Vol. 38. Applications of Mathematics (New York). Springer-Verlag, New York, 1998, pp. xvi+396.
- [Din+22] Z. Ding, S. Chen, Q. Li, and S. J. Wright. “Overparameterization of deep ResNet: zero loss and mean-field analysis.” In: *J. Mach. Learn. Res.* 23 (2022), Paper No. [48], 65.
- [Don06] D. L. Donoho. “Compressed sensing.” In: *IEEE Trans. Inform. Theory* 52.4 (2006), pp. 1289–1306.
- [DB05] M. Dorigo and C. Blum. “Ant colony optimization theory: A survey.” In: *Theoret. Comput. Sci.* 344.2-3 (2005), pp. 243–278.
- [DL09] P. Doukhan and G. Lang. “Evaluation for moments of a ratio with application to regression estimation.” In: *Bernoulli* 15.4 (2009), pp. 1259–1286.
- [Duc+15] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. “Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations.” In: *IEEE Trans. Inf. Theory* 61.5 (2015), pp. 2788–2806.
- [DHS11] J. Duchi, E. Hazan, and Y. Singer. “Adaptive subgradient methods for online learning and stochastic optimization.” In: *J. Mach. Learn. Res.* 12 (2011), pp. 2121–2159.
- [Dur+20] A. Durmus, A. Eberle, A. Guillin, and R. Zimmer. “An elementary approach to uniform in time propagation of chaos.” In: *Proc. Amer. Math. Soc.* 148.12 (2020), pp. 5387–5398.

- [DM17] A. Durmus and É. Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm.” In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587.
- [Dut+21] S. Dutta, T. Gautam, S. Chakrabarti, and T. Chakraborty. “Redesigning the Transformer Architecture with Insights from Multi-particle Dynamical Systems.” In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 5531–5544.
- [EHL19] W. E, J. Han, and Q. Li. “A mean-field optimal control formulation of deep learning.” In: *Res. Math. Sci.* 6.1 (2019), Paper No. 10, 41.
- [Eic21] G. Eichfelder. “Twenty years of continuous multiobjective optimization in the twenty-first century.” In: *EURO J. Comput. Optim.* 9 (2021), Paper No. 100014, 15.
- [ERY24] B. Engquist, K. Ren, and Y. Yang. “Adaptive state-dependent diffusion for derivative-free optimization.” In: *Communications on Applied Mathematics and Computation* (2024), pp. 1–29.
- [ERY22] B. Engquist, K. Ren, and Y. Yang. “An algebraically converging stochastic gradient descent algorithm for global optimization.” In: *arXiv preprint arXiv:2204.05923* (2022).
- [Erm75] D. L. Ermak. “A computer simulation of charged particles in solution. I. Technique and equilibrium properties.” In: *The Journal of Chemical Physics* 62.10 (1975), pp. 4189–4196.
- [FF22] X. Fernández-Real and A. Figalli. “The continuous formulation of shallow neural networks as wasserstein-type gradient flows.” In: *Analysis at Large: Dedicated to the Life and Work of Jean Bourgain*. Springer, 2022, pp. 29–57.
- [Fil+08] C. J. A. B. Filho, F. B. de Lima Neto, A. J. da Cunha Carneiro Lins, A. I. S. Nascimento, and M. P. Lima. “A novel search algorithm based on fish school behavior.” In: *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12-15 October 2008*. IEEE, 2008, pp. 2646–2651.
- [Fle01] R. Fletcher. *Practical methods of optimization*. second. Wiley-Interscience [John Wiley & Sons], New York, 2001, pp. xiv+436.
- [Fog00] D. B. Fogel. *Evolutionary computation. Toward a new philosophy of machine intelligence*. Second. IEEE Press, Piscataway, NJ, 2000, pp. xx+270.
- [For+22] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. “Anisotropic diffusion in consensus-based optimization on the sphere.” In: *SIAM J. Optim.* 32.3 (2022), pp. 1984–2012.

- [For+20] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. “Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit.” In: *Math. Models Methods Appl. Sci.* 30.14 (2020), pp. 2725–2751.
- [For+21] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. “Consensus-based optimization on the sphere: convergence to global minimizers and machine learning.” In: *J. Mach. Learn. Res.* 22 (2021), Paper No. 237, 55.
- [CBO-I] M. Fornasier, T. Klock, and K. Riedl. “Consensus-Based Optimization Methods Converge Globally.” In: *arXiv preprint arXiv:2103.15130* (2021).
- [CBO-II] M. Fornasier, T. Klock, and K. Riedl. “Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law.” In: *Applications of Evolutionary Computation - 25th European Conference, EvoApplications 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings*. Ed. by J. L. J. Laredo, J. I. Hidalgo, and K. O. Babaagba. Vol. 13224. Lecture Notes in Computer Science. Springer, 2022, pp. 738–754.
- [CBO-III] M. Fornasier, P. Richtárik, K. Riedl, and L. Sun. “Consensus-Based Optimization with Truncated Noise.” In: *Eur. J. Appl. Math. (special issue “From integro-differential models to data-oriented approaches for emergent phenomena”)* (accepted 2024, to appear).
- [FS14] M. Fornasier and F. Solombrino. “Mean-field optimal control.” In: *ESAIM Control Optim. Calc. Var.* 20.4 (2014), pp. 1123–1152.
- [FS24] M. Fornasier and L. Sun. “A PDE Framework of Consensus-Based Optimization for Objectives with Multiple Global Minimizers.” In: *arXiv preprint arXiv:2403.06662* (2024).
- [For+05] S. Fortunato, V. Latora, A. Pluchino, and A. Rapisarda. “Vector opinion dynamics in a bounded confidence consensus model.” In: *International Journal of Modern Physics C* 16.10 (2005), pp. 1535–1551.
- [FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013, pp. xviii+625.
- [GP06] M. Garavello and B. Piccoli. “Traffic flow on networks.” In: *AIMS Series on Applied Mathematics* 1 (2006). Conservation laws models, pp. xvi+243.
- [GRV23] G. Garrigos, L. Rosasco, and S. Villa. “Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry.” In: *Math. Program.* 198.1 (2023), pp. 937–996.
- [GS90] A. E. Gelfand and A. F. M. Smith. “Sampling-based approaches to calculating marginal densities.” In: *J. Amer. Statist. Assoc.* 85.410 (1990), pp. 398–409.
- [GM91] S. B. Gelfand and S. K. Mitter. “Recursive stochastic algorithms for global optimization in \mathbb{R}^d .” In: *SIAM J. Control Optim.* 29.5 (1991), pp. 999–1018.

- [GH86] S. Geman and C.-R. Hwang. “Diffusions for global optimization.” In: *SIAM J. Control Optim.* 24.5 (1986), pp. 1031–1043.
- [GP19] M. Gendreau and J.-Y. Potvin, eds. *Handbook of metaheuristics*. Third. Vol. 272. International Series in Operations Research & Management Science. Springer, Cham, 2019, pp. xx+604.
- [Ger23] N. J. Gerber. *Consensus-Based Optimization and Sampling: Propagation of Chaos via the Coupling Method*. 2023.
- [GHV23] N. J. Gerber, F. Hoffmann, and U. Vaes. “Mean-field limits for Consensus-Based Optimization and Sampling.” In: *arXiv preprint arXiv:2312.07373* (2023).
- [Ges+23a] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. “A mathematical perspective on Transformers.” In: *arXiv preprint arXiv:2312.10794* (2023).
- [Ges+23b] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. “The emergence of clusters in self-attention dynamics.” In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 57026–57037.
- [Gho+20] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran. “An Efficient Framework for Clustered Federated Learning.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 19586–19597.
- [Gib60] J. W. Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics*. Dover Publications, Inc., New York, 1960, pp. xviii+207.
- [GMW20] P. E. Gill, W. Murray, and M. H. Wright. *Practical optimization*. Vol. 81. Classics in Applied Mathematics. Reprint of the 1981 original [0634376]. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, [2020] ©2020, pp. xx+401.
- [Gla+10] T. Glasmachers, T. Schaul, Y. Sun, D. Wierstra, and J. Schmidhuber. “Exponential natural evolution strategies.” In: *Genetic and Evolutionary Computation Conference, GECCO 2010, Proceedings, Portland, Oregon, USA, July 7-11, 2010*. Ed. by M. Pelikan and J. Branke. ACM, 2010, pp. 393–400.
- [Goh+09] R. H. Gohary, Y. Huang, Z.-Q. Luo, and J.-S. Pang. “A generalized iterative water-filling algorithm for distributed power control in the presence of a jammer.” In: *IEEE Trans. Signal Process.* 57.7 (2009), pp. 2660–2674.
- [GGL12] J. Gómez-Serrano, C. Graham, and J.-Y. Le Boudec. “The bounded confidence model of opinion dynamics.” In: *Math. Models Methods Appl. Sci.* 22.2 (2012), pp. 1150007, 46.

- [Goo+20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. “Generative adversarial networks.” In: *Commun. ACM* 63.11 (2020), pp. 139–144.
- [GM97] C. Graham and S. Méléard. “Stochastic particle approximations for generalized Boltzmann models and convergence estimates.” In: *Ann. Probab.* 25.1 (1997), pp. 115–132.
- [Gra+23] S. Grassi, H. Huang, L. Pareschi, and J. Qiu. “Mean-field particle swarm optimization.” In: *Modeling and Simulation for Collective Dynamics*. World Scientific, 2023, pp. 127–193.
- [GP21] S. Grassi and L. Pareschi. “From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit.” In: *Math. Models Methods Appl. Sci.* 31.8 (2021), pp. 1625–1657.
- [HJK21] S.-Y. Ha, S. Jin, and D. Kim. “Convergence and error estimates for time-discrete consensus-based optimization algorithms.” In: *Numer. Math.* 147.2 (2021), pp. 255–282.
- [HJK20] S.-Y. Ha, S. Jin, and D. Kim. “Convergence of a first-order consensus-based global optimization algorithm.” In: *Math. Models Methods Appl. Sci.* 30.12 (2020), pp. 2417–2444.
- [Had02] J. Hadamard. “Sur les problèmes aux dérivées partielles et leur signification physique.” In: *Princeton university bulletin* (1902), pp. 49–52.
- [Han06] N. Hansen. “The CMA Evolution Strategy: A Comparing Review.” In: *Towards a New Evolutionary Computation - Advances in the Estimation of Distribution Algorithms*. Ed. by J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea. Vol. 192. Studies in Fuzziness and Soft Computing. Springer, 2006, pp. 75–102.
- [HO96] N. Hansen and A. Ostermeier. “Adapting Arbitrary Normal Mutation Distributions in Evolution Strategies: The Covariance Matrix Adaptation.” In: *Proceedings of 1996 IEEE International Conference on Evolutionary Computation, Nayoya University, Japan, May 20-22, 1996*. Ed. by T. Fukuda and T. Furuhashi. IEEE, 1996, pp. 312–317.
- [HO01] N. Hansen and A. Ostermeier. “Completely Derandomized Self-Adaptation in Evolution Strategies.” In: *Evol. Comput.* 9.2 (2001), pp. 159–195.
- [Has70] W. K. Hastings. “Monte Carlo sampling methods using Markov chains and their applications.” In: *Biometrika* 57.1 (1970), pp. 97–109.
- [Haz16] E. Hazan. “Introduction to Online Convex Optimization.” In: *Found. Trends Optim.* 2.3-4 (2016), pp. 157–325.
- [HWO23] H. Heaton, S. Wu Fung, and S. Osher. “Global solutions to nonconvex problems by evolution of Hamilton-Jacobi PDEs.” In: *Communications on Applied Mathematics and Computation* (2023), pp. 1–21.

- [HK02] R. Hegselmann and U. Krause. “Opinion dynamics and bounded confidence: models, analysis and simulation.” In: *J. Artif. Soc. Soc. Simul.* 5.3 (2002).
- [Hig01] D. J. Higham. “An algorithmic introduction to numerical simulation of stochastic differential equations.” In: *SIAM Rev.* 43.3 (2001), pp. 525–546.
- [Hoe+21] T. Hoeffler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste. “Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks.” In: *J. Mach. Learn. Res.* 22 (2021), Paper No. 241, 124.
- [Hol75] J. H. Holland. *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor, Mich., 1975, pp. ix+183.
- [HD18] S. A. Hollingsworth and R. O. Dror. “Molecular dynamics simulation for all.” In: *Neuron* 99.6 (2018), pp. 1129–1143.
- [Hua21] H. Huang. “A note on the mean-field limit for the particle swarm optimization.” In: *Appl. Math. Lett.* 117 (2021), Paper No. 107133, 9.
- [HQ22] H. Huang and J. Qiu. “On the mean-field limit for the consensus-based optimization.” In: *Math. Methods Appl. Sci.* 45.12 (2022), pp. 7814–7831.
- [CBO-SP] H. Huang, J. Qiu, and K. Riedl. “Consensus-Based Optimization for Saddle Point Problems.” In: *SIAM J. Control Optim.* 62.2 (2024), pp. 1093–1121.
- [PSO] H. Huang, J. Qiu, and K. Riedl. “On the Global Convergence of Particle Swarm Optimization Methods.” In: *Appl. Math. Optim.* 88.2 (2023), Paper No. 30, 44.
- [Hwa80] C.-R. Hwang. “Laplace’s method revisited: weak convergence of probability measures.” In: *Ann. Probab.* 8.6 (1980), pp. 1177–1182.
- [HM79] C. L. Hwang and A. S. M. Masud. *Multiple objective decision making—methods and applications*. Vol. 164. Lecture Notes in Economics and Mathematical Systems. A state-of-the-art survey, In collaboration with Sudhakar R. Paidy and Kwangsun Yoon. Springer-Verlag, Berlin-New York, 1979, pp. xii+351.
- [Isa12] V. Isaeva. “Self-organization in biological systems.” In: *Biology Bulletin* 39 (2012), pp. 110–118.
- [JW17] P.-E. Jabin and Z. Wang. “Mean field limit for stochastic particle systems.” In: *Active particles. Vol. 1. Advances in theory, models, and applications*. Model. Simul. Sci. Eng. Technol. Birkhäuser/Springer, Cham, 2017, pp. 379–402.
- [Jah04] J. Jahn. *Vector optimization. Theory, applications, and extensions*. Springer-Verlag, Berlin, 2004, pp. xiv+465.
- [JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. “The variational formulation of the Fokker-Planck equation.” In: *SIAM J. Math. Anal.* 29.1 (1998), pp. 1–17.

- [Kac56] M. Kac. “Foundations of kinetic theory.” In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. III*. Univ. California Press, Berkeley-Los Angeles, Calif., 1956, pp. 171–197.
- [Kai+21] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. A. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D’Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. “Advances and Open Problems in Federated Learning.” In: *Found. Trends Mach. Learn.* 14.1-2 (2021), pp. 1–210.
- [Kal+24] D. Kalise, E. Loayza-Romero, K. A. Morris, and Z. Zhong. “Multi-level Optimal Control with Neural Surrogate Models.” In: *arXiv preprint arXiv:2402.07763* (2024).
- [KST23] D. Kalise, A. Sharma, and M. V. Tretyakov. “Consensus-based optimization via jump-diffusion stochastic differential equations.” In: *Math. Models Methods Appl. Sci.* 33.2 (2023), pp. 289–339.
- [KNS16] H. Karimi, J. Nutini, and M. Schmidt. “Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition.” In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*. Ed. by P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken. Vol. 9851. Lecture Notes in Computer Science. Springer, 2016, pp. 795–811.
- [KP90] M. Karplus and G. A. Petsko. “Molecular dynamics simulations in biology.” In: *Nature* 347.6294 (1990), pp. 631–639.
- [Ken97] J. Kennedy. “The particle swarm: social adaptation of knowledge.” In: *Proceedings of 1997 IEEE International Conference on Evolutionary Computation*. IEEE, 1997, pp. 303–308.
- [KE95] J. Kennedy and R. Eberhart. “Particle swarm optimization.” In: *Proceedings of International Conference on Neural Networks (ICNN’95), Perth, WA, Australia, November 27 - December 1, 1995*. IEEE, 1995, pp. 1942–1948.
- [Kim+20] J. Kim, M. Kang, D. Kim, S. Ha, and I. Yang. “A Stochastic Consensus Method for Nonconvex Optimization on the Stiefel Manifold.” In: *59th IEEE Conference on Decision and Control, CDC 2020, Jeju Island, South Korea, December 14-18, 2020*. IEEE, 2020, pp. 1050–1057.

- [KB15] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization.” In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. 2015.
- [KGV83] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. “Optimization by simulated annealing.” In: *Science* 220.4598 (1983), pp. 671–680.
- [Kir84] S. Kirkpatrick. “Optimization by simulated annealing: quantitative studies.” In: *J. Statist. Phys.* 34.5-6 (1984), pp. 975–986.
- [KST24] K. Klamroth, M. Stiglmayr, and C. Totzeck. “Consensus-based optimization for multi-objective problems: a multi-swarm approach.” In: *Journal of Global Optimization* (2024), pp. 1–32.
- [KP92] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Vol. 23. Applications of Mathematics (New York). Springer-Verlag, Berlin, 1992, pp. xxxvi+632.
- [KHdS04] R. A. Krohling, F. Hoffmann, and L. dos Santos Coelho. “Co-evolutionary particle swarm optimization for min-max problems using Gaussian distribution.” In: *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2004, 19-23 June 2004, Portland, OR, USA*. IEEE, 2004, pp. 959–964.
- [LSS23] S. C.-H. Lam, J. Sirignano, and K. Spiliopoulos. “Kernel Limit of Recurrent Neural Networks Trained on Ergodic Data Sequences.” In: *arXiv preprint arXiv:2308.14555* (2023).
- [LPV02] E. C. Laskari, K. E. Parsopoulos, and M. N. Vrahatis. “Particle swarm optimization for minimax problems.” In: *Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002, Honolulu, HI, USA, May 12-17, 2002*. IEEE, 2002, pp. 1576–1581.
- [LL07] J.-M. Lasry and P.-L. Lions. “Mean field games.” In: *Jpn. J. Math.* 2.1 (2007), pp. 229–260.
- [LCB10] Y. LeCun, C. Cortes, and C. Burges. *MNIST handwritten digit database*. 2010.
- [Liu01] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New York, 2001, pp. xvi+343.
- [LDL13] Y. Liu, Y. Dai, and Z. Luo. “Max-Min Fairness Linear Transceiver Design for a Multi-User MIMO Interference Channel.” In: *IEEE Trans. Signal Process.* 61.9 (2013), pp. 2413–2423.
- [Lor63] E. N. Lorenz. “Deterministic nonperiodic flow.” In: *J. Atmospheric Sci.* 20.2 (1963), pp. 130–141.
- [LTZ22] J. Lu, E. Tadmor, and A. Zenginoglu. “Swarm-based gradient descent method for non-convex optimization.” In: *arXiv preprint arXiv:2211.17157* (2022).

- [Lu+19] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu. “Understanding and improving transformer from a multi-particle dynamic system point of view.” In: *arXiv preprint arXiv:1906.02762* (2019).
- [Mad+18a] D. Madras, E. Creager, T. Pitassi, and R. Zemel. “Learning Adversarially Fair and Transferable Representations.” In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 3384–3393.
- [Mad+18b] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards Deep Learning Models Resistant to Adversarial Attacks.” In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [Man63] B. Mandelbrot. “The Variation of Certain Speculative Prices.” In: *The Journal of Business* 36.4 (1963), pp. 394–419.
- [Már97] D. Márquez. “Convergence rates for annealing diffusion processes.” In: *Ann. Appl. Probab.* 7.4 (1997), pp. 1118–1139.
- [McK67] H. P. McKean Jr. “Propagation of chaos for a class of non-linear parabolic equations.” In: *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967)*. Vol. Session 7. Lecture Series in Differential Equations. Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Arlington, VA, 1967, pp. 41–57.
- [McM+17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 20–22 Apr 2017, pp. 1273–1282.
- [MMM19] S. Mei, T. Misiakiewicz, and A. Montanari. “Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit.” In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. PMLR, 25–28 Jun 2019, pp. 2388–2464.
- [MMN18] S. Mei, A. Montanari, and P.-M. Nguyen. “A mean field view of the landscape of two-layer neural networks.” In: *Proc. Natl. Acad. Sci. USA* 115.33 (2018), E7665–E7671.
- [Mil06] P. D. Miller. *Applied asymptotic analysis*. Vol. 75. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2006, pp. xvi+467.

- [MB11] E. Moulines and F. Bach. “Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning.” In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Vol. 24. Curran Associates, Inc., 2011.
- [MK87] K. G. Murty and S. N. Kabadi. “Some NP-complete problems in quadratic and nonlinear programming.” In: *Math. Programming* 39.2 (1987), pp. 117–129.
- [Mye91] R. B. Myerson. *Game theory. Analysis of conflict*. Harvard University Press, Cambridge, MA, 1991, pp. xvi+568.
- [Nas50] J. F. Nash Jr. “Equilibrium points in n -person games.” In: *Proc. Nat. Acad. Sci. U.S.A.* 36 (1950), pp. 48–49.
- [NNG19] I. Necoara, Y. Nesterov, and F. Glineur. “Linear convergence of first order methods for non-strongly convex optimization.” In: *Math. Program.* 175.1-2, Ser. A (2019), pp. 69–107.
- [NM65] J. A. Nelder and R. Mead. “A simplex method for function minimization.” In: *Comput. J.* 7.4 (1965), pp. 308–313.
- [NS17] Y. Nesterov and V. Spokoiny. “Random gradient-free minimization of convex functions.” In: *Found. Comput. Math.* 17.2 (2017), pp. 527–566.
- [New87] I. Newton. *Philosophiæ Naturalis Principia Mathematica*. Philos. Trans. R. Soc. A, 1687.
- [NP23] P.-M. Nguyen and H. T. Pham. “A rigorous framework for the mean field limit of multilayer neural networks.” In: *Math. Stat. Learn.* 6.3-4 (2023), pp. 201–357.
- [Nik22] V. Nikoghosyan. *Consensus-based solution of differential equations*. 2022.
- [Nik+22] K. Nikolakakis, F. Haddadpour, D. Kalogieras, and A. Karbasi. “Black-Box Generalization: Stability of Zeroth-Order Learning.” In: vol. 35. 2022, pp. 31525–31541.
- [Noc92] J. Nocedal. “Theory of algorithms for unconstrained optimization.” In: *Acta numerica, 1992*. Acta Numer. Cambridge Univ. Press, Cambridge, 1992, pp. 199–242.
- [NW06] J. Nocedal and S. J. Wright. *Numerical optimization*. Second. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006, pp. xxii+664.
- [Nou+19] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. “Solving a Class of Non-Convex Min-Max Games Using Iterative First Order Methods.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

- [Øks03] B. Øksendal. *Stochastic differential equations: An introduction with applications*. Sixth. Springer-Verlag, Berlin, 2003, pp. xxiv+360.
- [Oll+17] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. “Information-geometric optimization algorithms: a unifying picture via invariance principles.” In: *J. Mach. Learn. Res.* 18 (2017), Paper No. 18, 65.
- [Orl85] H. Orland. “Mean-field theory for optimization problems.” In: *Journal de Physique Lettres* 46.17 (1985), pp. 763–770.
- [OM98] E. Ozcan and C. K. Mohan. “Analysis of a simple particle swarm optimization system.” In: *Intelligent engineering systems through artificial neural networks* 8 (1998), pp. 253–258.
- [PSL11] B. K. Panigrahi, Y. Shi, and M.-H. Lim, eds. *Handbook of swarm intelligence*. Vol. 8. Adaptation, Learning, and Optimization. Concepts, principles and applications. Springer-Verlag, Berlin, 2011, pp. xii+543.
- [PŽŽ17] P. M. Pardalos, A. Žilinskas, and J. Žilinskas. *Non-convex multi-objective optimization*. Vol. 123. Springer Optimization and Its Applications. Springer, Cham, 2017, pp. xi+192.
- [Par24] L. Pareschi. “Optimization by linear kinetic equations and mean-field Langevin dynamics.” In: *arXiv preprint arXiv:2401.05553* (2024).
- [PE99] J. K. Parrish and L. Edelstein-Keshet. “Complexity, pattern, and evolutionary trade-offs in animal aggregation.” In: *Science* 284.5411 (1999), pp. 99–101.
- [Pas+19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.
- [Pel98] M. Pelletier. “Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing.” In: *Ann. Appl. Probab.* 8.1 (1998), pp. 10–44.
- [Pha+06] D. T. Pham, A. Ghanbarzadeh, E. Koç, S. Otri, S. Rahim, and M. Zaidi. “The bees algorithm—a novel tool for complex optimisation problems.” In: *Intelligent production machines and systems*. Elsevier, 2006, pp. 454–459.
- [Pin+17] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. “A consensus-based model for global optimization and its mean-field limit.” In: *Math. Models Methods Appl. Sci.* 27.1 (2017), pp. 183–204.
- [Pla99] E. Platen. “An introduction to numerical methods for stochastic differential equations.” In: *Acta numerica, 1999*. Vol. 8. Acta Numer. Cambridge Univ. Press, Cambridge, 1999, pp. 197–246.

- [Pol63] B. T. Polyak. “Gradient methods for minimizing functionals.” In: *Ž. Vychisl. Mat i Mat. Fiz.* 3 (1963), pp. 643–653.
- [Pol64] B. T. Polyak. “Some methods of speeding up the convergence of iterative methods.” In: *Ž. Vychisl. Mat i Mat. Fiz.* 4 (1964), pp. 791–803.
- [RN04] M. G. Rabbat and R. D. Nowak. “Distributed optimization in sensor networks.” In: *Proceedings of the Third International Symposium on Information Processing in Sensor Networks, IPSN 2004, Berkeley, California, USA, April 26-27, 2004*. Ed. by K. Ramchandran, J. Sztipanovits, J. C. Hou, and T. N. Pappas. ACM, 2004, pp. 20–27.
- [RT18] J. Rapin and O. Teytaud. *Nevergrad — A gradient-free optimization platform*. 2018.
- [Ras63] L. A. Rastrigin. “The convergence of the random search method in the extremal control of a many parameter system.” In: *Automaton & Remote Control* 24 (1963), pp. 1337–1342.
- [Raz+20] M. Razaviyayn, T. Huang, S. Lu, M. Nouiehed, M. Sanjabi, and M. Hong. “Nonconvex Min-Max Optimization: Applications, Challenges, and Recent Theoretical Advances.” In: *IEEE Signal Process. Mag.* 37.5 (2020), pp. 55–66.
- [RY99] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Third. Vol. 293. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999, pp. xiv+602.
- [Rey87] C. W. Reynolds. “Flocks, herds and schools: A distributed behavioral model.” In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, Anaheim, California, USA, July 27-31, 1987*. Ed. by M. C. Stone. ACM, 1987, pp. 25–34.
- [CBO-IV] K. Riedl. “Leveraging Memory Effects and Gradient Information in Consensus-Based Optimisation: On Global Convergence in Mean-Field Law.” In: *Eur. J. Appl. Math.* (accepted 2023, to appear), 32 pages.
- [CBO&GD] K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier. “Gradient is All You Need?” In: *arXiv preprint arXiv:2306.09778* (2023).
- [RC04] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Second. Springer Texts in Statistics. Springer-Verlag, New York, 2004, pp. xxx+645.
- [RT96] G. O. Roberts and R. L. Tweedie. “Exponential convergence of Langevin distributions and their discrete approximations.” In: *Bernoulli* 2.4 (1996), pp. 341–363.

- [RV22] G. M. Rotskoff and E. Vanden-Eijnden. “Trainability and accuracy of artificial neural networks: an interacting particle system approach.” In: *Comm. Pure Appl. Math.* 75.9 (2022), pp. 1889–1935.
- [RV18] G. Rotskoff and E. Vanden-Eijnden. “Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.
- [San17] F. Santambrogio. “{Euclidean, metric, and Wasserstein} gradient flows: an overview.” In: *Bull. Math. Sci.* 7.1 (2017), pp. 87–154.
- [SMS21] F. Sattler, K. Müller, and W. Samek. “Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints.” In: *IEEE Trans. Neural Networks Learn. Syst.* 32.8 (2021), pp. 3710–3722.
- [STW23] C. Schillings, C. Totzeck, and P. Wacker. “Ensemble-based gradient inference for particle methods in optimization and sampling.” In: *SIAM/ASA J. Uncertain. Quantif.* 11.3 (2023), pp. 757–787.
- [SW15] M. Schmitt and R. Wanka. “Particle swarm optimization almost surely finds local optima.” In: *Theoret. Comput. Sci.* 561.part A (2015), pp. 57–72.
- [Sch95] H.-P. Schwefel. *Evolution and optimum seeking*. Sixth-Generation Computer Technology Series. With 1 IBM PC floppy disk (3.5 inch; HD), A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1995, pp. x+444.
- [SR95] H. Schwefel and G. Rudolph. “Contemporary Evolution Strategies.” In: *Advances in Artificial Life, Third European Conference on Artificial Life, Granada, Spain, June 4-6, 1995, Proceedings*. Ed. by F. Morán, A. Moreno, J. J. M. Guervós, and P. Chacón. Vol. 929. Lecture Notes in Computer Science. Springer, 1995, pp. 893–907.
- [Sha17] O. Shamir. “An optimal algorithm for bandit and zero-order convex optimization with two-point feedback.” In: *J. Mach. Learn. Res.* 18 (2017), Paper No. 52, 11.
- [SE98] Y. Shi and R. Eberhart. “A modified particle swarm optimizer.” In: *1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360)*. IEEE, 1998, pp. 69–73.
- [SK02] Y. Shi and R. A. Krohling. “Co-evolutionary particle swarm optimization to solve min-max problems.” In: *Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002, Honolulu, HI, USA, May 12-17, 2002*. IEEE, 2002, pp. 1682–1687.

- [SS15] R. Shokri and V. Shmatikov. “Privacy-preserving deep learning.” In: *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*. IEEE, 2015, pp. 909–910.
- [SS22] J. Sirignano and K. Spiliopoulos. “Mean field analysis of deep neural networks.” In: *Math. Oper. Res.* 47.1 (2022), pp. 120–152.
- [SS20a] J. Sirignano and K. Spiliopoulos. “Mean field analysis of neural networks: a central limit theorem.” In: *Stochastic Process. Appl.* 130.3 (2020), pp. 1820–1852.
- [SS20b] J. Sirignano and K. Spiliopoulos. “Mean field analysis of neural networks: a law of large numbers.” In: *SIAM J. Appl. Math.* 80.2 (2020), pp. 725–752.
- [Sum10] D. J. Sumpter. *Collective animal behavior*. Princeton University Press, 2010.
- [Sun+09] Y. Sun, D. Wierstra, T. Schaul, and J. Schmidhuber. “Stochastic search using the natural gradient.” In: *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*. Ed. by A. P. Danyluk, L. Bottou, and M. L. Littman. Vol. 382. ACM International Conference Proceeding Series. ACM, 2009, pp. 1161–1168.
- [SB98] R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998.
- [Szn84] A.-S. Sznitman. “Nonlinear reflecting diffusion process, and the propagation of chaos and fluctuations associated.” In: *J. Funct. Anal.* 56.3 (1984), pp. 311–336.
- [Szn91] A.-S. Sznitman. “Topics in propagation of chaos.” In: *École d’Été de Probabilités de Saint-Flour XIX—1989*. Vol. 1464. Lecture Notes in Math. Springer, Berlin, 1991, pp. 165–251.
- [TZ24] E. Tadmor and A. Zenginoğlu. “Swarm-Based Optimization with Random Descent.” In: *Acta Appl. Math.* 190 (2024), Paper No. 2.
- [Tal09] E. Talbi. *Metaheuristics - From Design to Implementation*. Wiley, 2009.
- [TW20] C. Totzeck and M.-T. Wolfram. “Consensus-based global optimization with personal best.” In: *Math. Biosci. Eng.* 17.5 (2020), pp. 6026–6044.
- [TK13] M. Treiber and A. Kesting. “Traffic flow dynamics.” In: (2013). *Data, models and simulation*, Translated by Treiber and Christian Thiemann, pp. xiv+503.
- [van07] F. van den Bergh. “An analysis of particle swarm optimizers.” PhD thesis. University of Pretoria, 2007.
- [vdBE10] F. van den Bergh and A. P. Engelbrecht. “A convergence proof for the particle swarm optimiser.” In: *Fund. Inform.* 105.4 (2010), pp. 341–374.

- [Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need.” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.
- [Ver+21] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer. “A Survey on Distributed Machine Learning.” In: *ACM Comput. Surv.* 53.2 (2021), 30:1–30:33.
- [VZ12] T. Vicsek and A. Zafeiris. “Collective motion.” In: *Physics reports* 517.3-4 (2012), pp. 71–140.
- [Vli20] L. C. Vlieger. *Consensus-based Optimization on Graphs. Using an Interacting Particle System Approach*. 2020.
- [vNM07] J. von Neumann and O. Morgenstern. “Theory of games and economic behavior.” In: anniversary. With an introduction by Harold W. Kuhn and an afterword by Ariel Rubinstein. Princeton University Press, Princeton, NJ, 2007, pp. xxxii+739.
- [WTL18] D. Wang, D. Tan, and L. Liu. “Particle swarm optimization algorithm: an overview.” In: *Soft Comput.* 22.2 (2018), pp. 387–408.
- [Wib18] A. Wibisono. “Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem.” In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 2093–3027.
- [Wie+14] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. “Natural evolution strategies.” In: *J. Mach. Learn. Res.* 15 (2014), pp. 949–980.
- [Wie+08] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. “Natural Evolution Strategies.” In: *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2008, June 1-6, 2008, Hong Kong, China*. IEEE, 2008, pp. 3381–3387.
- [Wit11] C. Witt. “Theory of particle swarm optimization.” In: *Theory of randomized search heuristics*. Vol. 1. Ser. Theor. Comput. Sci. World Sci. Publ., Hackensack, NJ, 2011, pp. 197–223.
- [Xu+18] P. Xu, J. Chen, D. Zou, and Q. Gu. “Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.

- [XLY17] Y. Xu, Q. Lin, and T. Yang. “Adaptive SVRG Methods under Error Bound Conditions with Unknown Growth Parameter.” In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.
- [Yan+19] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson. “A survey of distributed optimization.” In: *Annu. Rev. Control.* 47 (2019), pp. 278–305.
- [Yan13] X. Yang. “Metaheuristic Optimization: Nature-Inspired Algorithms and Applications.” In: *Artificial Intelligence, Evolutionary Computing and Metaheuristics - In the Footsteps of Alan Turing*. Ed. by X. Yang. Vol. 427. Studies in Computational Intelligence. Springer, 2013, pp. 405–420.
- [Yan14] X.-S. Yang. *Nature-inspired optimization algorithms*. Elsevier, Inc., Amsterdam, 2014, pp. xii+263.
- [Yan+18] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. “Mean Field Multi-Agent Reinforcement Learning.” In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by J. G. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 5567–5576.
- [YY15] Q. Yuan and G. Yin. “Analyzing convergence and rates of convergence of particle swarm optimization algorithms using stochastic approximation methods.” In: *IEEE Trans. Automat. Control* 60.7 (2015), pp. 1760–1773.
- [ZWJ15] Y. Zhang, S. Wang, and G. Ji. “A comprehensive survey on particle swarm optimization algorithm and its applications.” In: *Math. Probl. Eng.* 931256 (2015), pp. 1–38.

Part II

**Publications and
Preprints**

Paper P1

Consensus-Based Optimization Methods Converge Globally

M. Fornasier, T. Klock, and K. Riedl
arXiv preprint arXiv:2103.15130 (2021)

Paper Summary of [CBO-I]³⁴

In the paper “Consensus-Based Optimization Methods Converge Globally,” we give the first holistic global convergence proof of CBO methods (2.2) and (2.10) with isotropic noise in (2.5) on the plane, thereby providing a novel analytical framework.

CBO, as proposed in [Pin+17], is a multi-agent metaheuristic derivative-free optimization method that can globally minimize nonconvex nonsmooth functions, i.e., solve problems of the form (2.1). Its design follows the guiding principles of metaheuristic algorithms, in particular particle swarm optimization [KE95; Ken97] and simulated annealing [AK89]. But, it is of much simpler nature in order to be amenable to theoretical analysis using ideas from statistical mechanics by taking a mean-field perspective and gaining an understanding of the macroscopic behavior of the dynamics [Car+18].

Based on an experimentally supported intuition, see [CBO-I, Figure 1], that, on average, CBO performs a gradient descent of the squared Euclidean distance to the global minimizer, we devise in [CBO-I] a novel technique for proving the convergence to the global minimizer in mean-field law for a rich class of objective functions [CBO-I, Theorem 3.7]. The result unveils internal mechanisms of CBO that are responsible for the success of the method. In particular, we prove that CBO performs a convexification of a very large class of optimization problems as the number of optimizing agents goes to infinity. Furthermore, we improve prior analyses by requiring minimal assumptions about the initialization of the method and by covering objectives that are merely locally Lipschitz continuous. As a core component of this analysis, we establish a quantitative nonasymptotic Laplace principle [CBO-I, Proposition 4.5], which may be of independent interest. From the result of CBO convergence in mean-field law, it becomes apparent that the hardness of any global optimization problem is necessarily encoded in the rate of the mean-field approximation, for which we provide a novel probabilistic quantitative estimate [CBO-I, Proposition 3.11]. The combination of the former results about the convergence in mean-field law and the quantitative mean-field approximation together with classical results of numerical approximation of SDEs allows to obtain probabilistic global convergence guarantees of the numerical CBO method [CBO-I, Theorem 3.8].

KR’s Contributions. Building upon ideas of MF to design and investigate a suitable Lyapunov functional for CBO, all authors collaborated on working out the technical details for proving the global convergence of the mean-field dynamics of CBO to a global minimizer, in particular, eventually identifying the Wasserstein-2 distance to be the correct quantity to study. TK wrote a first draft of the paper, together with KR, which was then refined by all authors. KR conducted the numerical experiments. At a later stage, KR developed and added the results about the mean-field approximation of CBO, consulting regularly with MF. This finally allowed to obtain a holistic convergence proof of the numerical CBO method in form of probabilistic global convergence guarantees. KR rewrote large parts of the paper, which was then proofread and refined by MF.

³⁴In this section, we follow [CBO-I, Abstract].

The following document is a reprint of

[CBO-I] M. Fornasier, T. Klock, and K. Riedl. “Consensus-Based Optimization Methods Converge Globally.” In: *arXiv preprint arXiv:2103.15130* (2021).

The permission to reprint and include the material is provided after the reprint.

Consensus-Based Optimization Methods Converge Globally

Massimo Fornasier^{*1,2,3}, Timo Klock^{†4,5} and Konstantin Riedl^{‡1,2}

¹*Technical University of Munich, School of Computation, Information and Technology,
Department of Mathematics, Munich, Germany*

²*Munich Center for Machine Learning, Munich, Germany*

³*Munich Data Science Institute, Munich, Germany*

⁴*Simula Research Laboratory, Department of Numerical Analysis and Scientific
Computing, Oslo, Norway*

⁵*University of San Diego, California, Department of Mathematics, San Diego, USA*

Abstract

In this paper we study consensus-based optimization (CBO), which is a multi-agent meta-heuristic derivative-free optimization method that can globally minimize nonconvex non-smooth functions and is amenable to theoretical analysis. Based on an experimentally supported intuition that, on average, CBO performs a gradient descent of the squared Euclidean distance to the global minimizer, we devise a novel technique for proving the convergence to the global minimizer in mean-field law for a rich class of objective functions. The result unveils internal mechanisms of CBO that are responsible for the success of the method. In particular, we prove that CBO performs a convexification of a large class of optimization problems as the number of optimizing agents goes to infinity. Furthermore, we improve prior analyses by requiring mild assumptions about the initialization of the method and by covering objectives that are merely locally Lipschitz continuous. As a core component of this analysis, we establish a quantitative nonasymptotic Laplace principle, which may be of independent interest. From the result of CBO convergence in mean-field law, it becomes apparent that the hardness of any global optimization problem is necessarily encoded in the rate of the mean-field approximation, for which we provide a novel probabilistic quantitative estimate. The combination of these results allows to obtain probabilistic global convergence guarantees of the numerical CBO method.

Keywords: global optimization, derivative-free optimization, nonsmoothness, nonconvexity, metaheuristics, consensus-based optimization, mean-field limit, Fokker-Planck equations

AMS subject classifications: 65K10, 90C26, 90C56, 35Q90, 35Q84

1 Introduction

A long-standing problem in applied mathematics is the global minimization of a potentially nonconvex nonsmooth cost function $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ and the search for an associated globally

*Email: massimo.fornasier@ma.tum.de

†Email: timo@simula.no

‡Email: konstantin.riedl@ma.tum.de

minimizing argument v^* . Throughout, we assume the unique existence of the minimizer v^* and denote its associated minimal value by

$$\underline{\mathcal{E}} := \mathcal{E}(v^*) = \inf_{v \in \mathbb{R}^d} \mathcal{E}(v).$$

The objective \mathcal{E} is supposed to be locally Lipschitz continuous and to satisfy a tractability condition of the form $\|v - v^*\|_2 \leq (\mathcal{E}(v) - \underline{\mathcal{E}})^\nu / \eta$ in a neighborhood of v^* , see Assumption A2 for the details. While computing $\underline{\mathcal{E}}$ or v^* are in general NP-hard problems under such conditions, several instances arising in real-world scenarios can, at least empirically, be solved within reasonable accuracy and moderate computational time. In the present work we are concerned with the class of derivative-free optimization algorithms, i.e., methods that are based exclusively on the evaluation of the objective function \mathcal{E} . Amongst them and achieving the state of the art on challenging problems such as the Traveling Salesman Problem, are so-called metaheuristics [1, 4, 5, 42, 55]. Metaheuristics orchestrate an interaction between local improvement procedures and global strategies, and combine deterministic and random decisions, to create a process capable of escaping from local optima and performing a robust search of the solution space. Examples include Random Search [54], Evolutionary Programming [24], the Metropolis-Hastings algorithm [33], Genetic Algorithms [35], Particle Swarm Optimization [42], and Simulated Annealing [1]. Despite their tremendous empirical success and widespread use in practice, many metaheuristics, due to their complexity, lack a proper mathematical foundation that could prove robust convergence to global minimizers under suitable assumptions. Nevertheless, for some of them, such as Random Search or Simulated Annealing, there exist probabilistic guarantees for global convergence, see, e.g., [36, 62]. While transferring some of the ideas of [62] to Particle Swarm Optimization allows to establish guaranteed convergence to global minima, the proof argument uses a computational time coinciding with the time necessary to examine every location in the search space [65].

Recently, the authors of [10, 52] have introduced consensus-based optimization (CBO) methods, which follow the guiding principles of metaheuristic algorithms, but are of much simpler nature and more amenable to theoretical analysis. Inspired by consensus dynamics and opinion formation, CBO methods use a finite number of agents V^1, \dots, V^N , which are formally stochastic processes, to explore the domain and to form a global consensus about the location of the minimizer v^* as time passes. The dynamics of the agents V^1, \dots, V^N are governed by two competing terms. A drift term drags each agent towards an instantaneous consensus point, denoted by v_α , which is computed as a weighted average of all agents' positions and serves as a momentaneous proxy for the global minimizer v^* . This term may be deactivated individually for an agent if its position improves upon the consensus point through modulating the drift by a function H approximating the Heaviside function. The second term is stochastic and randomly diffuses agents according to a scaled Brownian motion in \mathbb{R}^d , featuring the exploration of the energy landscape of the cost \mathcal{E} . Ideally, as result of the described drift-diffusion mechanism, the agents eventually achieve a near optimal global consensus, in the sense that the associated empirical measure

$$\widehat{\rho}_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{V_t^i} \quad (1)$$

converges to a Dirac delta $\delta_{\tilde{v}}$ at some $\tilde{v} \in \mathbb{R}^d$ close to v^* .

Let us now provide a formal description of the method. Given a time horizon $T > 0$ and a time discretization $t_0 = 0 < \Delta t < \dots < K\Delta t = T$ of $[0, T]$, we denote the location of agent i at time $k\Delta t$ by $V_{k\Delta t}^i$, $k = 0, \dots, K$. For user-specified parameters $\alpha, \lambda, \sigma > 0$, the time-discrete

evolution of the i -th agent is defined by the update rule

$$V_{(k+1)\Delta t}^i - V_{k\Delta t}^i = -\Delta t \lambda (V_{k\Delta t}^i - v_\alpha(\widehat{\rho}_{k\Delta t}^N)) H(\mathcal{E}(V_{k\Delta t}^i) - \mathcal{E}(v_\alpha(\widehat{\rho}_{k\Delta t}^N))) + \sigma \|V_{k\Delta t}^i - v_\alpha(\widehat{\rho}_{k\Delta t}^N)\|_2 B_{k\Delta t}^i, \quad (2)$$

$$V_0^i \sim \rho_0 \quad \text{for all } i = 1, \dots, N, \quad (3)$$

where $((B_{k\Delta t}^i)_{k=0, \dots, K-1})_{i=1, \dots, N}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}_d$. The system is complemented with independent initial data $(V_0^i)_{i=1, \dots, N}$, distributed according to a common initial law ρ_0 . Equation (2) originates from a simple Euler-Maruyama time discretization [34, 53] of the system of stochastic differential equations (SDEs)

$$dV_t^i = -\lambda (V_t^i - v_\alpha(\widehat{\rho}_t^N)) H(\mathcal{E}(V_t^i) - \mathcal{E}(v_\alpha(\widehat{\rho}_t^N))) dt + \sigma \|V_t^i - v_\alpha(\widehat{\rho}_t^N)\|_2 dB_t^i, \quad (4)$$

$$V_0^i \sim \rho_0 \quad \text{for all } i = 1, \dots, N, \quad (5)$$

where $((B_t^i)_{t \geq 0})_{i=1, \dots, N}$ are now independent standard Brownian motions in \mathbb{R}^d . As mentioned in the informal description above, the updates in the evolutions (2) and (4) consist of two terms, respectively. The first term is the drift towards the momentaneous consensus $v_\alpha(\widehat{\rho}_t^N)$, which is defined by

$$v_\alpha(\widehat{\rho}_t^N) := \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L^1(\widehat{\rho}_t^N)}} d\widehat{\rho}_t^N(v), \quad \text{with } \omega_\alpha(v) := \exp(-\alpha \mathcal{E}(v)). \quad (6)$$

Definition (6) is motivated by the well-known Laplace principle [21, 49, 52], which states that, for any absolutely continuous probability distribution ϱ on \mathbb{R}^d , we have

$$\lim_{\alpha \rightarrow \infty} \left(-\frac{1}{\alpha} \log \left(\int \omega_\alpha(v) d\varrho(v) \right) \right) = \inf_{v \in \text{supp}(\varrho)} \mathcal{E}(v). \quad (7)$$

Alternatively, we can also interpret (6) as an approximation of $\arg \min_{i=1, \dots, N} \mathcal{E}(V_t^i)$, which improves as $\alpha \rightarrow \infty$, provided the minimizer uniquely exists. The univariate function $H : \mathbb{R} \rightarrow [0, 1]$ appearing in the first term of (2) and (4) can be used to deactivate the drift term for agents V_t^i , whose objective is better than the one of the momentaneous consensus, i.e., for which $\mathcal{E}(V_t^i) < \mathcal{E}(v_\alpha(\widehat{\rho}_t^N))$, by setting $H(x) \approx \mathbb{1}_{x \geq 0}$. The most frequently studied choice however is $H \equiv 1$. The second term in (2) and (4) encodes the diffusion or exploration mechanism of the algorithm. Intuitively, scaling by $\|V_t^i - v_\alpha(\widehat{\rho}_t^N)\|_2$ encourages agents far from the consensus point to explore larger regions, whereas agents close to the consensus point try to enhance their position only locally. Furthermore, the scaling is essential to eventually deactivate the Brownian motion and to achieve consensus among the individual agents.

CBO methods have been considered and analyzed in several recent papers [8, 10–12, 16, 25–28, 40, 43, 64], even for optimization problems in high-dimensional and non-Euclidean settings, and using more sophisticated rules for the parameter choices α and σ inspired by Simulated Annealing [11, 26]. Moreover, several variants of the dynamics have been proposed, such as ones integrating memory mechanisms [57, 64] or others using jump-diffusion processes [40]. To make the method feasible and competitive for large-scale applications, in particular, for problems arising in machine learning, random mini-batch sampling techniques have been employed when evaluating the objective function or computing the consensus point. This significantly reduces the computational and communication complexity of CBO methods [11, 28] and further enables the parallelization of the algorithm by evolving disjoint subsets of particles independently for some time with separate consensus points, before aligning the dynamics through a global communication step. However, despite bearing interesting questions concerning the trade-off between parallel efficiency and performance when it comes to the relevance of communication

between the individual agents, this is a so far largely unexplored area for CBO. As an example for the applicability of CBO to such high-dimensional problems, we refer to [11, 28, 57] where the method is used for training a shallow and a convolutional neural network for image classification of the MNIST database of handwritten digits [44], to the recent paper [13] where CBO is used in the setting of clustered federated learning, to [57] where a compressed sensing problem is solved, or to the line of works [25–27] where (2) and (4) are adapted to the sphere \mathbb{S}^{d-1} achieving near state-of-the-art performance on a phase retrieval, a robust subspace detection problem and when robustly computing eigenfaces. Recently, also general constrained optimization problems have been tackled by CBO through the use of penalization techniques, which allow to cast the constrained problem into an unconstrained optimization task [8, 12].

As initially mentioned, CBO methods are motivated by the urge to develop a class of metaheuristic algorithms with provable guarantees, while preserving their capabilities of escaping local minima through global optimization mechanisms. The main theoretical interest focuses on understanding when consensus formation of $\widehat{\rho}_t^N \rightarrow \delta_{\tilde{v}}$ occurs, and on quantitatively bounding the associated errors $\mathcal{E}(\tilde{v}) - \mathcal{E}$ and $\|\tilde{v} - v^*\|_2$. A theoretical analysis of the dynamics can either be done on the microscopic systems (2) or (4), as for instance in [31, 32], or, as in [10, 52], by analyzing the macroscopic behavior of the agent density through a mean-field limit associated with the particle-based dynamics (4), given, for initial data $\bar{V}_0 \sim \rho_0$, by

$$d\bar{V}_t = -\lambda (\bar{V}_t - v_\alpha(\rho_t)) H(\mathcal{E}(\bar{V}_t) - \mathcal{E}(v_\alpha(\rho_t))) dt + \sigma \|\bar{V}_t - v_\alpha(\rho_t)\|_2 dB_t, \quad (8)$$

where $\rho_t = \text{Law}(\bar{V}_t)$. The weak convergence of the microscopic system (4) to the mean-field limit (8), or, more precisely, of the empirical measure $\widehat{\rho}_t^N$ to ρ_t as $N \rightarrow \infty$, has been shown recently in [37], see also Remark 2 for additional details. This legitimates to analyze (8) in lieu of (4). The measure $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ with $\rho_t = \rho(t) = \text{Law}(\bar{V}_t)$ satisfies the nonlinear nonlocal Fokker-Planck equation

$$\partial_t \rho_t = \lambda \text{div}((v - v_\alpha(\rho_t)) H(\mathcal{E}(v) - \mathcal{E}(v_\alpha(\rho_t))) \rho_t) + \frac{\sigma^2}{2} \Delta(\|v - v_\alpha(\rho_t)\|_2^2 \rho_t) \quad (9)$$

in a weak sense (see Definition 5). Leveraging this partial differential equation (PDE), the authors of [10, 52] analyze the large time behavior of the particle density $t \mapsto \rho_t$ instead of the microscopic systems (2) and (4). Studying the mean-field limit (8) or (9) allows for agile deterministic calculus tools and typically leads to stronger theoretical results, which characterize the average agent behavior through the evolution of ρ . This analysis perspective is justified by the mean-field approximation, which quantifies the convergence of the microscopic system (4) to the mean-field limit (8) as the number of agents grows. We discuss results about the mean-field approximation in Remark 2 and make it rigorous in Proposition 16. Hence, in view of its validity and as already done in the preceding works [10, 52], in the first part of the paper we concentrate on establishing convergence in mean-field law for (4), as defined in Definition 1 below. That is, we analyze the mean-field dynamics (8) and (9) in place of the interacting particle system (4). Afterwards, by combining the mean-field approximation with convergence in mean-field law, we close the paper with a global convergence result for the numerical method (2).

Definition 1 (Convergence in mean-field law). *Let $F, G : \mathcal{P}(\mathbb{R}^d) \otimes \mathbb{R}^d \rightarrow \mathbb{R}^d$ be two functions and consider for $i = 1, \dots, N$ the SDEs expressed in Itô's form as*

$$dV_t^i = F(\widehat{\rho}_t^N, V_t^i) dt + G(\widehat{\rho}_t^N, V_t^i) dB_t^i, \quad \text{where } \widehat{\rho}_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{V_t^i}, \text{ and } V_0^i \sim \rho_0.$$

We say that this SDE system converges in mean-field law to $\tilde{v} \in \mathbb{R}^d$ if all solutions of

$$d\bar{V}_t = F(\rho_t, \bar{V}_t) dt + G(\rho_t, \bar{V}_t) dB_t, \quad \text{where } \rho_t = \text{Law}(\bar{V}_t), \text{ and } \bar{V}_0 \sim \rho_0,$$

satisfy $\lim_{t \rightarrow \infty} W_p(\rho_t, \delta_{\tilde{v}}) = 0$ for some Wasserstein- p distance W_p , $p \geq 1$.

Colloquially speaking, an interacting multi-particle system is said to converge *in mean-field law*, if the associated mean-field dynamics converges.

Remark 2 (Mean-field approximation). The definition of convergence in mean-field law as given in Definition 1 is justified as follows: As the number of agents N in the interacting particle system (4) tends to infinity, one expects that, for any particle V^i , the individual influence of any other particle disperses. This results in an averaged influence of the ensemble rather than an interacting nature of the system, and allows to describe the dynamics in the large-particle limit by the law ρ of the mono-particle process (8). This phenomenon is known as the mean-field approximation. More formally, as $N \rightarrow \infty$, we expect the empirical measure $\widehat{\rho}_t^N$ to converge in law to ρ_t for almost every t , see [39, Definition 1]. The classical way to establish such mean-field approximation is to prove, by means of the coupling method, propagation of chaos [47, 63], as implied for instance by

$$\max_{i=1, \dots, N} \sup_{t \in [0, T]} \mathbb{E} \|V_t^i - \bar{V}_t^i\|_2^2 \leq CN^{-1}, \quad (10)$$

where \bar{V}^i denote N i.i.d. copies of the mean-field dynamics (8), which are coupled to the processes V^i by choosing the same initial conditions as well as Brownian motion paths, see, e.g., the recent review [14, 15]. Despite being of fundamental numerical interest (since when combined with the convergence in mean-field law it allows to establish convergence of the interacting particle system itself), a quantitative result about the mean-field approximation of CBO as in (10) has been left as a difficult and open problem in [10, Remark 3.3] due to a lack of global Lipschitz continuity of the drift and diffusion terms, which impedes the application of McKean's theorem [15, Theorem 3.1].

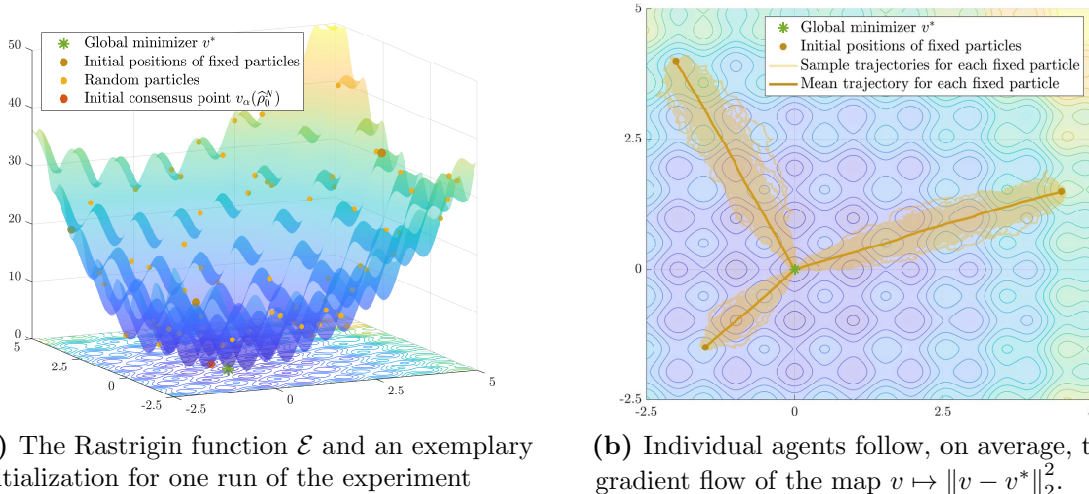
However, the present work as well as three recent works, which we outline in what follows, are shedding light on this issue. By employing a compactness argument in the path space, the authors of [37] show that the empirical random particle measure $\widehat{\rho}^N$ associated with the dynamics (4) converges in distribution to the deterministic particle distribution $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ satisfying (9). In particular, their result is valid for unbounded functions \mathcal{E} considered also in our work. While this does not allow for obtaining a quantitative convergence rate with respect to the number of particles N as in (10), it closes the mean-field limit gap qualitatively. A desired quantitative result has been established recently in [25, Theorem 3.1 and Remark 3.1] for a variant of the microscopic system (4) supported on a compact hypersurface Γ . In [25] the weak convergence of the variant of (4) to the corresponding mean-field limit is established in the sense that for all $\phi \in \mathcal{C}_b^1(\mathbb{R}^d)$ it holds

$$\sup_{t \in [0, T]} \mathbb{E} \left[|\langle \widehat{\rho}_t^N, \phi \rangle - \langle \rho_t, \phi \rangle|^2 \right] \leq \frac{C}{N} \|\phi\|_{\mathcal{C}^1(\mathbb{R}^d)}^2 \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

The obtained convergence rate reads CN^{-1} with C depending in particular on

$$C_\alpha := \exp(\alpha(\sup_{v \in \Gamma} \mathcal{E}(v) - \inf_{v \in \Gamma} \mathcal{E}(v))).$$

Their proof is based on the aforementioned coupling method and, by exploiting the inherent compactness of the dynamics due to its confinement to Γ , allows to derive a bound of the form (10). Leveraging the techniques from [25] and the boundedness of moments established in [10, Lemma 3.4], we provide in Proposition 16 below a result of the form (10) on the plane \mathbb{R}^d which holds with high probability. A more refined analysis conducted recently by the authors of [29], which adapts Sznitman's classical argument for the proof of McKean's theorem with the intention of allowing for coefficients which are not globally Lipschitz, even yields a non-probabilistic mean-field approximation of the form (10) in the pathwise sense, requiring in comparison merely a higher moment bound $\rho_0 \in \mathcal{P}_6(\mathbb{R}^d)$ of the initial measure, see [29, Theorem 2.6].



(a) The Rastrigin function \mathcal{E} and an exemplary initialization for one run of the experiment

(b) Individual agents follow, on average, the gradient flow of the map $v \mapsto \|v - v^*\|_2^2$.

Figure 1: An illustration of the internal mechanisms of CBO. We perform 100 runs of the CBO algorithm (2)–(3), with parameters $\Delta t = 0.01$, $\alpha = 10^{15}$, $\lambda = 1$ and $\sigma = 0.1$, and $N = 32000$ agents initialized according to $\rho_0 = \mathcal{N}((8, 8), 20)$. In addition, we add three individual agents with starting locations $(-2, 4)$, $(-1.5, -1.5)$ and $(4.5, 1.5)$ to the set of agents in each run as shown in (a), and depict each of their 100 trajectories as well as their mean trajectory in yellow color in (b). With the (mean) trajectories being rather straight lines, we observe that the individual agents take a straight path from their initial positions to the global minimizer v^* and, in particular, disregard the local landscape of the objective function \mathcal{E} . The trajectories of the individual agents become more concentrated as the overall number of agents N grows.

Such quantitative mean-field approximation results substantiate the focus of the first part of this work on the analysis of the macroscopic mean-field dynamics (8) and (9). Nevertheless, as a consequence thereof, we return to the analysis of the numerical scheme (2) and its global convergence in Section 3.3.

Contributions. In view of the versatility, simplicity, and efficiency of CBO methods, a theoretical understanding of the finite particle-based system (4) and the mean-field limit (8) is of great interest. In this work, we unveil the surprising phenomenon that, in the mean-field limit, for a rich class of objectives \mathcal{E} , the individual agents of the CBO dynamics follow the gradient flow associated with the function $v \mapsto \|v - v^*\|_2^2$, on average over all realizations of Brownian motion paths, see Figure 1. Interestingly, this gradient flow is independent of the underlying energy landscape of \mathcal{E} . In other words, CBO performs a canonical convexification of a large class of optimization problems as the number of optimizing agents N goes to infinity. Based on these observations and justified by the mean-field approximation, first of all we develop a novel proof framework for showing the convergence of the CBO dynamics in mean-field law to the global minimizer v^* for a rich class of objectives. While previous analyses in [10, 31, 32] required restrictive concentration conditions about the initial measure ρ_0 and \mathcal{C}^2 regularity of the objective, we derive results that are valid under mild assumptions about ρ_0 and local Lipschitz continuity of \mathcal{E} . We explain the key differences of this work with respect to previous work in detail in Section 2 and further showcase the benefits of the proposed analysis by a numerical example. These findings reveal that the hardness of any global optimization problem is necessarily encoded in the rate of the mean-field approximation as $N \rightarrow \infty$. Secondly, in consideration of its central significance with regards to the computational complexity of the numerical scheme (2) we establish a novel probabilistic quantitative result about the convergence of the interacting particle system (4) to the corresponding mean-field limit (9), which is a result of independent interest. By combining these two results, the convergence in mean-field law on

the one hand, and the quantitative mean-field approximation on the other, we provide the first, and so far unique, holistic convergence proof of CBO on the plane, enabling to quantify the optimization capability of the numerical CBO algorithm (2) in terms of the used parameters. The utilized proof technique may be used as a blueprint for proving global convergence for other recent adaptations of the CBO dynamics, see, e.g., [8, 11, 26–28, 40], as well as other metaheuristics such as the renowned Particle Swarm Optimization, which is related to CBO through a zero-inertia limit, see, e.g., [20, 30, 38]. While the present paper has foundational and theoretical nature and aims at completely clarifying the convergence of the numerical scheme (2) with a detailed analysis, we do not include extensive numerical experiments. For numerical evidence that CBO does solve difficult optimizations also in high dimensions without necessarily incurring in the curse of dimensionality, the reader may want to consult previous work such as [11, 13, 16, 26–28, 57].

Remark 3. Employing stochasticity and leveraging collaboration between multiple agents to empirically and provably achieve global convergence of numerical algorithms and to avoid convergence to local minima, is not just of particular relevance when it comes to the efficiency and success of zero-order methods, but also an emerging paradigm in the field of gradient-based optimization, see, e.g., [18, 22, 46]. Recent work [58] even suggests a connection between the worlds of derivative-free and gradient-based methods. Similar guiding principles are present also in sampling methods, such as Langevin sampling [17, 18, 23, 59] or Stein Variational Gradient Descent [45], which are designed to generate samples from an unknown target distribution.

A promising way to gain a theoretical understanding of the behavior of these classes of algorithms is by taking a mean-field perspective, i.e., by analyzing the dynamics, as the number of particles goes to infinity, through an associated PDE. This typically involves Polyak-Lojasiewicz-like conditions [41] or certain families of log-Sobolev inequalities [18] on the objective function \mathcal{E} , which are more restrictive than the assumptions under which the statements of this work hold. For a recent analysis of the mean-field Langevin dynamics we refer to [18] and references therein.

Lately and conceptually similar to the convexification of a highly nonconvex problem observed in this work, taking a mean-field perspective has allowed the authors of [19, 48, 60, 61] to explain the generalization capabilities of over-parameterized neural networks. By leveraging that the mean-field description (w.r.t. the number of neurons) of the SGD learning dynamics is captured by a nonlinear PDE, which admits a gradient flow structure on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, these works show that original complexities of the loss landscape are alleviated. Together with a quantification of the fluctuations of the empirical neuron distribution around this mean-field limit (i.e., a mean-field approximation), convergence results are derived for SGD for sufficiently large networks with optimal generalization error. These results, however, are structurally different from the ones obtained in this paper for CBO. In particular, the individual particles in [19, 48, 60, 61] are associated with the different neurons of a two-layer or deep neural network and the objective function is a specific empirical risk, which itself is subject to the mean-field limit and gains convexity as the number of neurons tends to infinity. In contrast, in our setting each particle itself is a competitor for minimization of a general fixed nonconvex objective function \mathcal{E} and the convexification of the problem emerges from the CBO dynamics when its mean-field limit behavior is studied. For this reason, the two resulting mean-field limits are different.

Let us further point out that, as far as the community could understand up to now, the Fokker-Planck equation (9) describing the mean-field behavior of CBO cannot be understood as a gradient flow of any energy on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Yet, and perhaps surprisingly, the analysis of our present paper shows that the Wasserstein-2 distance from the global minimizer is the correct Lyapunov functional to be analyzed.

1.1 Organization

In Section 2 we first discuss state-of-the-art global convergence results for CBO methods with a detailed account of the utilized proof technique, including potential weaknesses. The second part of Section 2 then motivates an alternative proof strategy and explains how it can remedy the weaknesses of prior proofs under minimalistic assumptions. In Section 3 we first provide additional details about the well-posedness of the macroscopic SDE (8), respectively, the Fokker-Planck equation (9), before presenting and discussing the main result about the convergence of the dynamics (8) and (9) to the global minimizer in mean-field law. In order to demonstrate the relevance of such statement in establishing a holistic convergence guarantee for the numerical scheme (2), we conclude the section with a probabilistic quantitative result about the mean-field approximation. Sections 4 and 5 comprise the proof details of the convergence result in mean-field law and the result about the mean-field approximation, respectively. Section 6 concludes the paper.

1.2 Notation

Euclidean balls are denoted as $B_r(u) := \{v \in \mathbb{R}^d : \|v - u\|_2 \leq r\}$. For the space of continuous functions $f : X \rightarrow Y$ we write $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$ and a suitable topological space Y . For an open set $X \subset \mathbb{R}^n$ and for $Y = \mathbb{R}^m$ the spaces $\mathcal{C}_c^k(X, Y)$ and $\mathcal{C}_b^k(X, Y)$ contain functions $f \in \mathcal{C}(X, Y)$ that are k -times continuously differentiable and have compact support or are bounded, respectively. We omit Y in the real-valued case. The operators ∇ and Δ denote the gradient and Laplace operator of a function on \mathbb{R}^d . The main objects of study are laws of stochastic processes, $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, where the set $\mathcal{P}(\mathbb{R}^d)$ contains all Borel probability measures over \mathbb{R}^d . With $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ we refer to a snapshot of such law at time t . In case we refer to some fixed distribution, we write ϱ . Measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite p -th moment $\int \|v\|_2^p d\varrho(v)$ are collected in $\mathcal{P}_p(\mathbb{R}^d)$. For any $1 \leq p < \infty$, W_p denotes the Wasserstein- p distance between two Borel probability measures $\varrho_1, \varrho_2 \in \mathcal{P}_p(\mathbb{R}^d)$, see, e.g., [2]. $\mathbb{E}(\varrho)$ denotes the expectation of a probability measure ϱ .

2 Blueprints for the analysis of CBO methods

In this section we provide intuitive descriptions of two approaches to the analysis of the convergence of CBO methods to global minimizers. We first recall [10], and related works [31, 32], which prove convergence as a consequence of a monotonous decay of the variance of ρ_t and by employing the asymptotic Laplace principle (7). This proof strategy incurs a restrictive condition about the parameters α, λ, σ and the initial configuration ρ_0 , which implies that a small optimization gap $\mathcal{E}(\tilde{v}) - \mathcal{E}(v^*)$ can only be achieved for initial configurations ρ_0 already well-concentrated near the optimizer v^* . We then motivate an alternative proof idea to remedy this weakness based on the intuition that ρ_t monotonically minimizes the squared Euclidean distance to the global minimizer v^* .

2.1 State of the art: variance-based convergence analysis

We now recall the blueprint proof strategy from [10], which has been adapted in other works, e.g., [26, 31, 32], to prove consensus formation and convergence to the global minimum.

A successful application of the CBO framework underlies the premise that the induced particle density ρ_t converges to a Dirac delta $\delta_{\tilde{v}}$ for some \tilde{v} close to v^* . The analysis in [10] proves this under certain assumptions by first showing that ρ_t converges to a Dirac delta around *some* $\tilde{v} \in \mathbb{R}^d$ and then concluding $\tilde{v} \approx v^*$ in a subsequent step. Regarding the first step, the authors of [10] study the variance of ρ_t , defined as $\text{Var}(\rho_t) := \frac{1}{2} \int \|v - \mathbb{E}(\rho_t)\|_2^2 d\rho_t(v)$,

where $\mathbb{E}(\rho_t) := \int v d\rho_t(v)$, and show that $\text{Var}(\rho_t)$ decays exponentially fast in t under a well-preparedness assumption about the initial condition ρ_0 . More precisely, in [10, Section 4.1] the authors use Itô's lemma to derive for the time-evolution of $\text{Var}(\rho_t)$ the expression

$$\frac{d}{dt} \text{Var}(\rho_t) = - (2\lambda - d\sigma^2) \text{Var}(\rho_t) + \frac{d\sigma^2}{2} \|\mathbb{E}(\rho_t) - v_\alpha(\rho_t)\|_2^2. \quad (11)$$

For parameter choices $2\lambda > d\sigma^2$, the first term in (11) is negative and one could *almost* apply Grönwall's inequality to obtain the asserted exponential decay of $\text{Var}(\rho_t)$. However, the second term can be problematic and the main difficulty is to control the distance $\|\mathbb{E}(\rho_t) - v_\alpha(\rho_t)\|_2$ between the mean and the weighted mean. For $\alpha \rightarrow 0$ the weight function $\omega_\alpha(v) = \exp(-\alpha\mathcal{E}(v))$ associated with $v_\alpha(\rho_t)$ converges to 1 pointwise and consequently $v_\alpha(\rho_t) \rightarrow \mathbb{E}(\rho_t)$. However, the second proof step, explained below, reveals that the crucial regime is $\alpha \gg 1$. In this case $v_\alpha(\rho_t)$ can be arbitrarily far from $\mathbb{E}(\rho_t)$ if we do not dispose of additional knowledge about the probability measure ρ_t . To restrict the set of probability measures ρ_t that need to be considered when bounding $\|\mathbb{E}(\rho_t) - v_\alpha(\rho_t)\|_2$, the authors of [10] compromise to assume that the initial distribution ρ_0 satisfies the well-preparedness assumptions

$$\alpha e^{-2\alpha\mathcal{E}}(\sigma^2 + 2\lambda) < 3/8 \quad \text{and} \quad 2\lambda \|\omega_\alpha\|_{L_1(\rho_0)}^2 - \text{Var}(\rho_0) - 2d\sigma^2 \|\omega_\alpha\|_{L_1(\rho_0)} e^{-\alpha\mathcal{E}} \geq 0. \quad (12)$$

Since ρ_t evolves from ρ_0 according to the Fokker-Planck equation (9), these conditions restrict ρ_t and allow for bounding $\|\mathbb{E}(\rho_t) - v_\alpha(\rho_t)\|_2$ by a suitable multiple of $\text{Var}(\rho_t)$. The exponential decay of $\text{Var}(\rho_t)$ then follows from (11) after applying Grönwall's inequality, see [10, Theorem 4.1]. Furthermore, the conditions in (12) also allow for proving convergence of ρ_t to a stationary Dirac delta at $\tilde{v} \in \mathbb{R}^d$.

Given convergence to a Dirac at \tilde{v} , in a second step it is shown $\mathcal{E}(\tilde{v}) \approx \mathcal{E}(v^*)$. In order to prove this approximation, one first deduces that for any $\varepsilon > 0$, there exists $\alpha \gg 1$ such that for all $t \geq 0$ it holds $-\frac{1}{\alpha} \log(\|\omega_\alpha\|_{L_1(\rho_t)}) \leq -\frac{1}{\alpha} \log(\|\omega_\alpha\|_{L_1(\rho_0)}) + \frac{\varepsilon}{2}$. This involves deriving a lower bound for the evolution $\frac{d}{dt} \|\omega_\alpha\|_{L_1(\rho_t)}$ for sufficiently large $\alpha > 0$ as done in [10, Lemma 4.1], which is then combined with the formerly proven exponentially decaying variance, see [10, Proof of Theorem 4.2]. Then, by recognizing that the Laplace principle (7) implies the existence of some $\alpha \gg 1$ with

$$-\frac{1}{\alpha} \log(\|\omega_\alpha\|_{L_1(\rho_0)}) - \underline{\mathcal{E}} < \frac{\varepsilon}{2}, \quad (13)$$

and by establishing the convergence $\|\omega_\alpha\|_{L_1(\rho_t)} \rightarrow \exp(-\alpha\mathcal{E}(\tilde{v}))$ as $t \rightarrow \infty$, one obtains the desired result $\mathcal{E}(\tilde{v}) - \underline{\mathcal{E}} < \varepsilon$ in the limit $t \rightarrow \infty$, see [10, Lemma 4.2]. The gap $\mathcal{E}(\tilde{v}) - \underline{\mathcal{E}}$ can be tightened by increasing α , but it is impossible to establish an explicit relation $\alpha = \alpha(\varepsilon)$ due to the use of the asymptotic Laplace principle.

This proof sketch unveils a tension on the role of the parameter α . Namely, the second step requires large $\alpha = \alpha(\varepsilon)$ to achieve $\mathcal{E}(\tilde{v}) - \underline{\mathcal{E}} < \varepsilon$. In fact, $\alpha(\varepsilon)$ may grow uncontrollably as we decrease the accuracy ε . The first step, however, requires the conditions in (12) which, in the most optimistic case, where $\sigma = 0$, imply

$$\text{Var}(\rho_0) \leq \frac{3}{8\alpha} \left(\int \exp(-\alpha(\mathcal{E}(v) - \underline{\mathcal{E}})) d\rho_0(v) \right)^2. \quad (14)$$

Therefore, ρ_0 needs to be increasingly concentrated as α increases, and should ideally be supported on sets where $\mathcal{E}(v) \approx \underline{\mathcal{E}}$. Designing such distribution ρ_0 in practice seems impossible in the absence of a good initial guess for v^* . In particular, we cannot expect (14) to hold for generic choices such as a uniform distribution on a compact set.

We add that the works [31, 32] conduct a similarly flavored analysis for the fully time-discretized microscopic system (2), with some differences in the details. They first show an

exponentially decaying variance under mild assumptions about λ and σ , but provided that the same Brownian motion is used for all agents, i.e., $(B_{k\Delta t}^i)_{k=1,\dots,K} = (B_{k\Delta t})_{k=1,\dots,K}$ for all $i = 1, \dots, N$. Such a choice leads to a less explorative dynamics, but it simplifies the consensus formation analysis. For proving $\mathcal{E}(\bar{v}) \approx \underline{\mathcal{E}}$ however, the authors again require an initial configuration ρ_0 that satisfies a technical concentration condition like (13), see for instance [32, Remark 3.1].

2.2 Alternative approach: CBO minimizes the squared distance to v^*

The approach described in the previous section might suggest that CBO only converges locally, which is in fact not what is observed in practice. Instead, global optimization is actually expected. To remedy the locality requirements of the variance-based analysis, let us now sketch and motivate an alternative proof idea. By averaging out the randomness associated with different realizations of Brownian motion paths, the macroscopic time-continuous SDE (8), in the case $H \equiv 1$, becomes

$$\frac{d}{dt} \mathbb{E}[\bar{V}_t | \bar{V}_0] = -\lambda \mathbb{E}[(\bar{V}_t - v_\alpha(\rho_t)) | \bar{V}_0] = -\lambda \mathbb{E}[(\bar{V}_t - v^*) | \bar{V}_0] + \lambda(v_\alpha(\rho_t) - v^*). \quad (15)$$

Furthermore, if \mathcal{E} is locally Lipschitz continuous and satisfies the coercivity condition

$$\|v - v^*\|_2 \leq \frac{1}{\eta} (\mathcal{E}(v) - \mathcal{E}(v^*))^\nu = \frac{1}{\eta} (\mathcal{E}(v) - \underline{\mathcal{E}})^\nu, \quad \text{for all } v \in \mathbb{R}^d, \quad (16)$$

and for some $\eta > 0$ and $\nu \in (0, \infty)$, the second term on the right-hand side of (15) can be made arbitrarily small for sufficiently large α , i.e., $v_\alpha(\rho_t) \approx v^*$ (more details follow below). In this case, the average dynamics of \bar{V}_t is well-approximated by

$$\frac{d}{dt} \mathbb{E}[\bar{V}_t | \bar{V}_0] \approx -\lambda \mathbb{E}[(\bar{V}_t - v^*) | \bar{V}_0], \quad (17)$$

which corresponds to the gradient flow of $v \mapsto \|v - v^*\|_2^2$ with rate 2λ . In other words, each individual agent essentially performs a gradient-descent of $v \mapsto \|v - v^*\|_2^2$ on average over all realizations of Brownian motion paths. Figure 1b visualizes this phenomenon for three isolated agents on the Rastrigin function in two dimensions.

Inspired by this observation, our proof strategy is to show that CBO methods successively minimize the energy functional $\mathcal{V} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}_{\geq 0}$, given by

$$\mathcal{V}(\rho_t) := \frac{1}{2} \int \|v - v^*\|_2^2 d\rho_t(v). \quad (18)$$

Note that this functional essentially coincides with the Wasserstein distance in the sense that $W_2^2(\rho_t, \delta_{v^*}) = 2\mathcal{V}(\rho_t)$. Therefore $\mathcal{V}(\rho_t) \rightarrow 0$ in particular implies that ρ_t converges weakly to δ_{v^*} , see [2, Chapter 7].

This novel approach does not suffer a tension on the parameter α like the variance-based analysis from the previous section. Roughly speaking (see Lemma 17 for details), $\mathcal{V}(\rho_t)$ follows an evolution similar to (11), with $\text{Var}(\rho_t)$ being replaced by $\mathcal{V}(\rho_t)$. However, we can now bound $\int \|v - v_\alpha(\rho_t)\|_2^2 d\rho_t(v) \leq 4\mathcal{V}(\rho_t) + 2\|v_\alpha(\rho_t) - v^*\|_2^2$, so that it just remains to control the second term. In comparison to bounding $\|v_\alpha(\rho_t) - \mathbb{E}(\rho_t)\|_2^2$ in terms of $\text{Var}(\rho_t)$ for the variance-based analysis, this requires to bound $\|v_\alpha(\rho_t) - v^*\|_2^2$ in terms of $\mathcal{V}(\rho_t)$. Fortunately, this is a much easier task: the Laplace principle generally asserts $\|v_\alpha(\rho_t) - v^*\|_2 \rightarrow 0$ under (16) as $\alpha \rightarrow \infty$ and we can even establish (see Proposition 21 for details) the quantitative estimate

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \frac{(2Lr)^\nu}{\eta} + \frac{\exp(-\alpha Lr)}{\varrho(B_r(v^*))} \int \|v - v^*\|_2 d\varrho(v)$$

for an arbitrary probability measure ρ and assuming that \mathcal{E} is L -Lipschitz in a ball of radius $r > 0$. This allows to estimate $\|v_\alpha(\rho_t) - v^*\|_2^2$ in terms of $\mathcal{V}(\rho_t)$ as desired.

Finally, we note that $\mathcal{V}(\rho_t)$ majorizes $\text{Var}(\rho_t)$ because $u \mapsto \frac{1}{2} \int \|v - u\|_2^2 d\rho_t(v)$ is minimized by the expectation $\mathbb{E}(\rho_t)$. This relation may be a source of concern, as it shows that proving $\mathcal{V}(\rho_t) \rightarrow 0$ implies $\text{Var}(\rho_t) \rightarrow 0$. We emphasize however that this does not imply a majorization for the corresponding time derivatives. In fact, Example 4 suggests that $\mathcal{V}(\rho_t)$ can decay exponentially while $\text{Var}(\rho_t)$ increases initially.

Example 4. We consider the Rastrigin function $\mathcal{E}(v) = v^2 + 2.5(1 - \cos(2\pi v))$ with global minimum at $v^* = 0$ and various local minima, see Figure 2a. For different initial configurations $\rho_0 = \mathcal{N}(\mu, 0.8)$ with $\mu \in \{1, 2, 3, 4\}$, we evolve the discretized system (2) using $N = 320000$ agents, discrete time step size $\Delta t = 0.01$ and parameters $\alpha = 10^{15}$ (i.e., the consensus point is the argmin of the agents), $\lambda = 1$ and $\sigma = 0.5$. By considering different means from $\mu = 1$ to $\mu = 4$, we push the global minimizer v^* into the tails of the initial configuration ρ_0 . Figure 2b shows that the decreasing initial probability mass around v^* eventually causes the variance $\text{Var}(\widehat{\rho}_t^N)$ (dashed lines) to increase in the beginning of the dynamics. In contrast, $\mathcal{V}(\widehat{\rho}_t^N)$ always decays exponentially fast with convergence speed $(2\lambda - d\sigma^2)$, independently of the initial condition ρ_0 . From a theoretical perspective, this means proving global convergence using a variance-based analysis as in Section 2.1 must require assumptions about ρ_0 such as Condition (14), whereas using $\mathcal{V}(\rho_t)$ does not suffer from this issue. The convergence speed $(2\lambda - d\sigma^2)$ coincides with the result in Theorem 12.

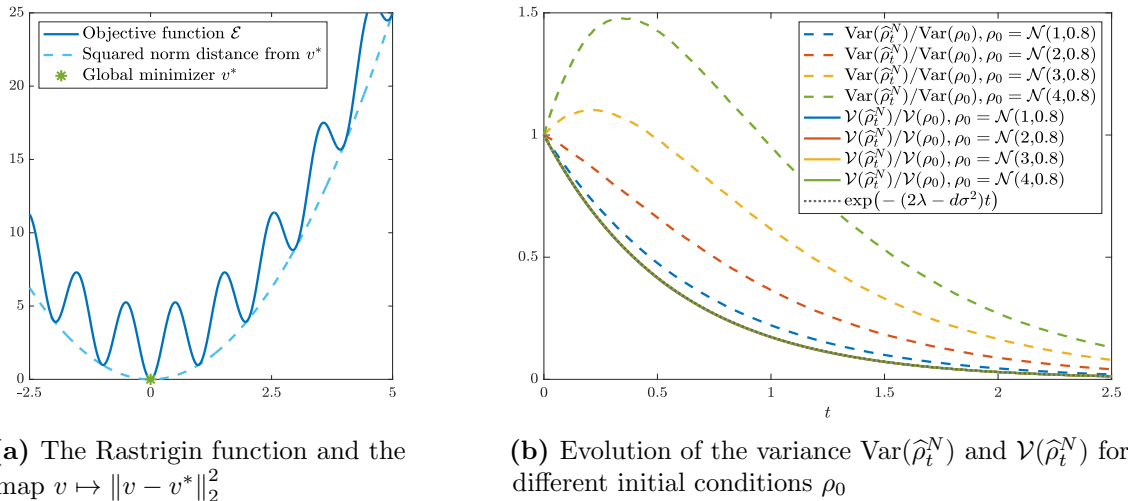


Figure 2: (a) The Rastrigin function as objective function \mathcal{E} and the squared Euclidean distance from v^* . (b) The evolution of the variance $\text{Var}(\widehat{\rho}_t^N)$ and the functional $\mathcal{V}(\widehat{\rho}_t^N)$ for different initial conditions $\rho_0 = \mathcal{N}(\mu, 0.8)$ with $\mu \in \{1, 2, 3, 4\}$. The measure $\widehat{\rho}_t^N$ is the empirical agent density that is evolved using (2) with $N = 320000$ agents, discrete time step size $\Delta t = 0.01$ and parameters $\alpha = 10^{15}$, $\lambda = 1$ and $\sigma = 0.5$. As we move the mean of the initial configuration ρ_0 away from the global optimizer $v^* = 0$, and thereby push v^* into the tails of ρ_0 , $\text{Var}(\widehat{\rho}_t^N)$ increases in the starting phase of the dynamics. $\mathcal{V}(\widehat{\rho}_t^N)$ on the other hand always decreases exponentially at a rate $(2\lambda - d\sigma^2)$, independently of the initial condition ρ_0 .

3 Global convergence of consensus-based optimization

In the first part of this section we recite and extend well-posedness results about the nonlinear macroscopic SDE (8), respectively, the associated Fokker-Planck equation (9). At the beginning of the second part we introduce the class of studied objective functions, which is followed by the

presentation of the main result about the convergence of the dynamics (8) and (9) to the global minimizer in mean-field law. In the final part we then highlight the relevance of this result by presenting a holistic convergence proof of the numerical scheme (2) to the global minimizer. This combines the latter statement with a probabilistic quantitative result about the mean-field approximation.

3.1 Definition of weak solutions and well-posedness

We begin by rigorously defining weak solutions of the Fokker-Planck equation (9).

Definition 5. Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$, $T > 0$. We say $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ satisfies the Fokker-Planck equation (9) with initial condition ρ_0 in the weak sense in the time interval $[0, T]$, if we have for all $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and all $t \in (0, T)$

$$\begin{aligned} \frac{d}{dt} \int \phi(v) d\rho_t(v) &= -\lambda \int H(\mathcal{E}(v) - \mathcal{E}(v_\alpha(\rho_t))) \langle v - v_\alpha(\rho_t), \nabla \phi(v) \rangle d\rho_t(v) \\ &\quad + \frac{\sigma^2}{2} \int \|v - v_\alpha(\rho_t)\|_2^2 \Delta \phi(v) d\rho_t(v) \end{aligned} \quad (19)$$

and $\lim_{t \rightarrow 0} \rho_t = \rho_0$ pointwise.

If the cutoff function H in the dynamics (8) is inactive, i.e., satisfies $H \equiv 1$, the authors of [10] prove the following well-posedness result.

Theorem 6 ([10, Theorems 3.1, 3.2]). Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Let $H \equiv 1$ and consider $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ with $\underline{\mathcal{E}} > -\infty$, which, for constants $C_1, C_2 > 0$, satisfies

$$|\mathcal{E}(v) - \mathcal{E}(w)| \leq C_1(\|v\|_2 + \|w\|_2) \|v - w\|_2, \quad \text{for all } v, w \in \mathbb{R}^d, \quad (20)$$

$$\mathcal{E}(v) - \underline{\mathcal{E}} \leq C_2(1 + \|v\|_2^2), \quad \text{for all } v \in \mathbb{R}^d. \quad (21)$$

If in addition, either $\sup_{v \in \mathbb{R}^d} \mathcal{E}(v) < \infty$, or \mathcal{E} satisfies for some constants $C_3, C_4 > 0$

$$\mathcal{E}(v) - \underline{\mathcal{E}} \geq C_3 \|v\|_2^2, \quad \text{for all } \|v\|_2 \geq C_4, \quad (22)$$

then there exists a unique nonlinear process $\bar{V} \in \mathcal{C}([0, T], \mathbb{R}^d)$ satisfying (8) in the strong sense. The associated law $\rho = \text{Law}(\bar{V})$ has regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ and is a weak solution to the Fokker-Planck equation (9).

Remark 7. The regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ stated in Theorem 6, and also obtained in Theorem 8 below, is a consequence of the regularity of the initial condition $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Despite not indicated explicitly in [10, Theorems 3.1, 3.2], it follows from their proofs. In particular, it allows for extending the test function space $\mathcal{C}_c^\infty(\mathbb{R}^d)$ in Definition 5. Namely, if $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ solves (9) in the weak sense, Identity (19) holds for all $\phi \in \mathcal{C}^2(\mathbb{R}^d)$ with (i) $\sup_{v \in \mathbb{R}^d} |\Delta \phi(v)| < \infty$, and (ii) $\|\nabla \phi(v)\|_2 \leq C(1 + \|v\|_2)$ for some $C > 0$ and for all $v \in \mathbb{R}^d$. We denote the corresponding function space by $\mathcal{C}_*^2(\mathbb{R}^d)$.

Under minor modifications of the proof for Theorem 6, we can extend the existence of solutions to an active Lipschitz-continuous cutoff function H .

Theorem 8. Let $H \neq 1$ be L_H -Lipschitz continuous. Then, under the assumptions of Theorem 6, there exists a nonlinear process $\bar{V} \in \mathcal{C}([0, T], \mathbb{R}^d)$ satisfying (8) in the strong sense. The associated law $\rho = \text{Law}(\bar{V})$ has regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ and is a weak solution to the Fokker-Planck equation (9).

3.2 Global convergence in mean-field law

We now present the main result about global convergence in mean-field law for objectives satisfying the following.

Definition 9 (Assumptions). *Throughout we are interested in objective functions $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$, for which*

A1 *there exists $v^* \in \mathbb{R}^d$ such that $\mathcal{E}(v^*) = \inf_{v \in \mathbb{R}^d} \mathcal{E}(v) =: \underline{\mathcal{E}}$, and*

A2 *there exist $\mathcal{E}_\infty, R_0, \eta > 0$, and $\nu \in (0, \infty)$ such that*

$$\|v - v^*\|_2 \leq (\mathcal{E}(v) - \underline{\mathcal{E}})^\nu / \eta \quad \text{for all } v \in B_{R_0}(v^*), \quad (23)$$

$$\mathcal{E}(v) - \underline{\mathcal{E}} > \mathcal{E}_\infty \quad \text{for all } v \in (B_{R_0}(v^*))^c. \quad (24)$$

Furthermore, for the case $H \neq 1$, we additionally require that \mathcal{E} fulfills a local Lipschitz continuity-like condition, i.e.,

A3 *there exist $L_\mathcal{E} > 0$ and $\gamma \geq 0$ such that*

$$\mathcal{E}(v) - \underline{\mathcal{E}} \leq L_\mathcal{E}(1 + \|v - v^*\|_2^\gamma) \|v - v^*\|_2 \quad \text{for all } v \in \mathbb{R}^d. \quad (25)$$

Remark 10. The analyses in [10] and related works require $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ and an additional boundedness assumptions on the Laplacian $\Delta\mathcal{E}$. We relax these regularity requirements and use the conditions in Definition 9 on \mathcal{E} instead.

Assumption A1 just states that the continuous objective \mathcal{E} attains its infimum $\underline{\mathcal{E}}$ at some $v^* \in \mathbb{R}^d$. The continuity itself can be further relaxed at the cost of additional technical details because it is only required in a small neighborhood of v^* .

Assumption A2 should be interpreted as a tractability condition of the landscape of \mathcal{E} around v^* and in the farfield. The first part, Equation (23), describes the local coercivity of \mathcal{E} , which implies that there is a unique minimizer v^* on $B_{R_0}(v^*)$ and that \mathcal{E} grows like $v \mapsto \|v - v^*\|_2^{1/\nu}$. This condition is also known as the inverse continuity condition from [26], as a quadratic growth condition in the case $\nu = 1/2$ from [3, 50], or as the Hölderian error bound condition in the case $\nu \in (0, 1]$ [6]. In [50, Theorem 4] and [41, Theorem 2] many equivalent or stronger conditions are identified to imply Equation (23) globally on \mathbb{R}^d . Furthermore, in [26, 66], (23) is shown to hold globally for objectives related to various machine learning problems. The second part of A2, Equation (24), describes the behavior of \mathcal{E} in the farfield and prevents $\mathcal{E}(v) \approx \underline{\mathcal{E}}$ for some $v \in \mathbb{R}^d$ far away from v^* . We introduce it for the purpose of covering functions that tend to a constant just above \mathcal{E}_∞ as $\|v\|_2 \rightarrow \infty$, because such functions do not satisfy the growth condition (23) globally. However, whenever (23) holds globally, we take $R_0 = \infty$, i.e., $B_{R_0}(v^*) = \mathbb{R}^d$ and (24) is void. We also note that (23) and (24) imply the uniqueness of the global minimizer v^* on \mathbb{R}^d .

Finally, to cover the active cutoff case $H \neq 1$, we additionally require A3. The condition is weaker than local Lipschitz-continuity on any compact ball around v^* , with Lipschitz constant growing with the size of the ball.

Example 11. A prototypical objective function that satisfies the assumptions in Definition 9 is the Rastrigin function, which we already discussed in Example 4, see also Figures 1a and 2a. In particular, it satisfies (23) globally with $\nu = 1/2$.

We are now ready to state the main result. The proof is deferred to Section 4.

Theorem 12. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ be such that $v^* \in \text{supp}(\rho_0)$. Define $\mathcal{V}(\rho_t)$ as given in (18). Fix any $\varepsilon \in (0, \mathcal{V}(\rho_0))$ and $\vartheta \in (0, 1)$, choose parameters $\lambda, \sigma > 0$ with $2\lambda > d\sigma^2$, and define the time horizon*

$$T^* := \frac{1}{(1 - \vartheta)(2\lambda - d\sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right). \quad (26)$$

Then there exists $\alpha_0 > 0$, depending (among problem dependent quantities) on ε and ϑ , such that for all $\alpha > \alpha_0$, if $\rho \in \mathcal{C}([0, T^], \mathcal{P}_4(\mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (9) on the time interval $[0, T^*]$ with initial condition ρ_0 , we have*

$$\mathcal{V}(\rho_T) = \varepsilon \quad \text{with} \quad T \in \left[\frac{1 - \vartheta}{(1 + \vartheta/2)} T^*, T^* \right]. \quad (27)$$

Furthermore, on the time interval $[0, T]$, $\mathcal{V}(\rho_t)$ decays at least exponentially fast. More precisely, for all $t \in [0, T]$, it holds

$$W_2^2(\rho_t, \delta_{v^*}) = 2\mathcal{V}(\rho_t) \leq 2\mathcal{V}(\rho_0) \exp \left(-(1 - \vartheta)(2\lambda - d\sigma^2)t \right). \quad (28)$$

If \mathcal{E} additionally satisfies A3, the same conclusion holds for any $H : \mathbb{R}^d \rightarrow [0, 1]$ that satisfies $H(x) = 1$ whenever $x \geq 0$.

The assumption $v^* \in \text{supp}(\rho_0)$ about the initial configuration ρ_0 is not really a restriction, as it would anyhow hold immediately for ρ_t for any $t > 0$ in view of the diffusive character of the dynamics (9), see Remark 23. Additionally, as we clarify in the next section, this condition does neither mean nor require that, for finite particle approximations, some particle needs to be in the vicinity of the minimizer v^* at time $t = 0$. It is actually sufficient that the empirical measure $\widehat{\rho}_t^N$ weakly approximates the law ρ_t uniformly in time. We rigorously explain this mechanism in Section 3.3.

A lower bound on the rate of convergence in (28) is $(1 - \vartheta)(2\lambda - d\sigma^2)$, which can be made arbitrarily close to the numerically observed rate $(2\lambda - d\sigma^2)$ (see, e.g., Figure 2b) at the cost of taking $\alpha \rightarrow \infty$ to allow for $\vartheta \rightarrow 0$. The condition $2\lambda > d\sigma^2$ is necessary, both in theory and practice, to avoid overwhelming the dynamics by the random exploration term. The dependency on d can be eased by replacing the isotropic Brownian motion in the dynamics with an anisotropic one [11, 28].

3.3 Global convergence in probability

To stress the relevance of the main result of this paper, Theorem 12, we now show how Estimate (28) plays a fundamental role in establishing a quantitative convergence result for the numerical scheme (2) to the global minimizer v^* . By paying the price of having a probabilistic statement about the convergence of CBO as in Theorem 13, we gain provable polynomial complexity. For simplicity, we present the results of this section for the case of an inactive cutoff function, i.e., $H \equiv 1$.

Theorem 13. *Fix $\varepsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Then, under the assumptions of Theorem 12 and Proposition 16, and with $K := T/\Delta t$, where T is as in (27), the iterations $((V_{k\Delta t}^i)_{k=0, \dots, K})_{i=1, \dots, N}$ generated by the numerical scheme (2) converge in probability to v^* . More precisely, the empirical mean of the final iterations fulfills*

$$\left\| \frac{1}{N} \sum_{i=1}^N V_{K\Delta t}^i - v^* \right\|_2^2 \leq \varepsilon_{\text{total}} \quad (29)$$

with probability larger than $1 - (\delta + \varepsilon_{\text{total}}^{-1}(6C_{\text{NA}}(\Delta t)^{2m} + 3C_{\text{MFA}}N^{-1} + 12\varepsilon))$. Here, m denotes the order of accuracy of the numerical scheme (for the Euler-Maruyama scheme $m = 1/2$)

and ε is the error from Theorem 12. Moreover, besides problem-dependent constants, $C_{\text{NA}} > 0$ depends linearly on the dimension d and the number of particles N , exponentially on the time horizon T , and on δ^{-1} ; $C_{\text{MFA}} > 0$ depends exponentially on the parameters α , λ and σ , on T , and on δ^{-1} .

Let us briefly discuss in the following remark the computational complexity of the numerical scheme (2) together with some implementational aspects which allow to reduce the overall runtime of the algorithm in practice.

Remark 14 (Computational complexity). To achieve Estimate (29) with probability of at least $(1 - 2\delta)$, the implementable CBO scheme (2) has to be run using $N \geq 9C_{\text{MFA}}/(\delta\varepsilon_{\text{total}})$ agents and with time step size $\Delta t \leq \sqrt[2m]{\delta\varepsilon_{\text{total}}/(18C_{\text{NA}})}$ for

$$K \geq \frac{1}{(1 - \vartheta)(2\lambda - d\sigma^2)} \frac{1}{\Delta t} \log \left(\frac{36\mathcal{V}(\rho_0)}{\delta\varepsilon_{\text{total}}} \right)$$

iterations. Here, the parameter dependence of C_{NA} and C_{MFA} is as described in Theorem 13. The computational complexity (counted in terms of the number of evaluations of the objective \mathcal{E}) of the CBO method is therefore given by $\mathcal{O}(KN)$.

When working in the setting of large-scale applications arising, for instance, in machine learning and signal processing (therefore, with \mathcal{E} being expensive to compute), several considerations allow to reduce the overall runtime of the algorithm (2) and thereby make the method feasible and more competitive. First of all, it may be recommendable to leverage that the evaluations of the objective function \mathcal{E} for each of the N particles can be performed in parallel. Furthermore, random mini-batch sampling ideas as proposed in [11, 28] may be employed when evaluating the objective function and/or computing the consensus point. I.e., at each time step, \mathcal{E} is evaluated only on a random subset of the available data, and v_α is computed only from a subset of the N particles. Besides immediately reducing the computational and communication complexity of CBO methods, such ideas motivate communication-efficient parallelization of the algorithm by evolving disjoint subsets of particles independently for some time with separate consensus points, before aligning the dynamics through a global communication step. This, however, is so far largely unexplored, both from a theoretical and practical point of view. Lastly, taking inspiration from genetic algorithms, a variance-based particle reduction technique as suggested in [26] may be used to reduce the number of optimizing agents (and therefore the required evaluations of \mathcal{E}) during the algorithm in case concentration of the particles is observed.

The proof of Theorem 13, which we report below, combines our main result about the convergence in mean-field law, a quantitative mean-field approximation and classical results of numerical approximation of SDEs. To this end, we establish in what follows the result about the quantitative mean-field approximation on a restricted set of bounded processes. For this purpose, let us introduce the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ over which all considered stochastic processes get their realizations, and define a subset Ω_M of Ω of suitably bounded processes according to

$$\Omega_M := \left\{ \omega \in \Omega : \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max \left\{ \|V_t^i(\omega)\|_2^4, \|\bar{V}_t^i(\omega)\|_2^4 \right\} \leq M \right\}.$$

Throughout this section, $M > 0$ denotes a constant which we shall adjust at the end of the proof of Theorem 13. Before stating the mean-field approximation result, Proposition 16, let us estimate the measure of the set Ω_M in Lemma 15. The proofs of both statements are deferred to Section 5.

Lemma 15. *Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N \in \mathbb{N}$ be fixed. Moreover, let $((V_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote the strong solution to system (4) and let $((\bar{V}_t^i)_{t \geq 0})_{i=1, \dots, N}$ be N independent copies of*

the strong solution to the mean-field dynamics (8). Then, under the assumptions of Theorem 6, for any $M > 0$ we have

$$\mathbb{P}(\Omega_M) = \mathbb{P}\left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max\{\|V_t^i\|_2^4, \|\bar{V}_t^i\|_2^4\} \leq M\right) \geq 1 - \frac{2K}{M}, \quad (30)$$

where $K = K(\lambda, \sigma, d, T, b_1, b_2)$ is a constant, which is in particular independent of N . Here, b_1 and b_2 denote the problem-dependent constants from [10, Lemma 3.3].

Lemma 15 proves that the processes are bounded with high probability uniformly in time. Therefore, by restricting the analysis to Ω_M , we can obtain the following quantitative mean-field approximation result by proving pointwise propagation of chaos through the coupling method [14, 15] using a synchronous coupling between the stochastic processes V^i and \bar{V}^i , see, e.g., [14, Section 4.1.2].

Proposition 16. *Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and let $N \in \mathbb{N}$ be fixed. Moreover, let $((V_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote the strong solution to system (4) and let $((\bar{V}_t^i)_{t \geq 0})_{i=1, \dots, N}$ be N independent copies of the strong solution to the mean-field dynamics (8). Further consider valid the assumptions of Theorem 6. If $(V_t^i)_{t \geq 0}$ and $(\bar{V}_t^i)_{t \geq 0}$ share the initial data as well as the Brownian motion paths $(B_t^i)_{t \geq 0}$ for all $i = 1, \dots, N$, then we have*

$$\max_{i=1, \dots, N} \sup_{t \in [0, T]} \mathbb{E}[\|V_t^i - \bar{V}_t^i\|_2^2 \mid \Omega_M] \leq C_{\text{MFA}} N^{-1} \quad (31)$$

with $C_{\text{MFA}} = C_{\text{MFA}}(\alpha, \lambda, \sigma, T, C_1, C_2, M, K, \mathcal{M}_2, b_1, b_2)$, where K is as in Lemma 15 and \mathcal{M}_2 denotes a second-order moment bound of ρ .

A quantitative mean-field approximation was left as an open problem in [10, Remark 3.2] due to a lack of global Lipschitz continuity of the SDE coefficients and approached since then in several steps, see Remark 2. While the restriction to bounded processes, which reflects the typical behavior in view of Lemma 15, already allows to obtain an estimate of the type (31), which is sufficient to prove convergence in probability in what follows, the recent work [29] improves (31) by firstly showing a non-probabilistic mean-field approximation, i.e., removing the necessity of conditioning on the set Ω_M as done in (31), and secondly by obtaining a pathwise estimate, see [29, Theorem 2.6]. Hence, in the light of [29], the role of the constant M can be regarded as merely an auxiliary technical tool.

Equipped with Lemma 15 and Proposition 16, we are now able to prove Theorem 13.

Proof of Theorem 13. We have the error decomposition

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N V_{K\Delta t}^i - v^* \right\|_2^2 \mid \Omega_M \right] &\leq 3 \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (V_{K\Delta t}^i - V_T^i) \right\|_2^2 \mid \Omega_M \right]}_{\leq C_{\text{NA}}(\Delta t)^{2m} \text{ by applying classical convergence results for numerical schemes for SDEs [53]} \\ &+ 3 \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (V_T^i - \bar{V}_T^i) \right\|_2^2 \mid \Omega_M \right]}_{\leq C_{\text{MFA}} N^{-1} \text{ using the quantitative mean-field approximation in form of Proposition 16}} + \frac{3}{1-\delta} \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_T^i - v^* \right\|_2^2 \right]}_{\leq \mathbb{E} \|\bar{V}_T^1 - v^*\|_2^2 \leq 2\mathcal{V}(\rho_T) \leq 2\varepsilon \text{ by means of Theorem 12}} \\ &\leq 6C_{\text{NA}}(\Delta t)^{2m} + 3C_{\text{MFA}} N^{-1} + 12\varepsilon \end{aligned} \quad (32)$$

dividing the overall error into an approximation error of the numerical scheme, the mean-field approximation error and the optimization error in the mean-field limit.

Denoting now by $K_{\varepsilon_{\text{total}}}^N \subset \Omega$ the set, where (29) does not hold, we can estimate

$$\begin{aligned} \mathbb{P}(K_{\varepsilon_{\text{total}}}^N) &= \mathbb{P}(K_{\varepsilon_{\text{total}}}^N \cap \Omega_M) + \mathbb{P}(K_{\varepsilon_{\text{total}}}^N \cap \Omega_M^c) \leq \mathbb{P}(K_{\varepsilon_{\text{total}}}^N \mid \Omega_M) \mathbb{P}(\Omega_M) + \mathbb{P}(\Omega_M^c) \\ &\leq \varepsilon_{\text{total}}^{-1} (6C_{\text{NA}}(\Delta t)^{2m} + 3C_{\text{MFA}}N^{-1} + 12\varepsilon) + \delta, \end{aligned}$$

where in the last step we employ Markov's inequality together with (32) to bound the first term. For the second it suffices to choose the M from (30) large enough. \square

As a consequence of Theorem 12, the hardness of any optimization problem is necessarily encoded in the mean-field approximation. Proposition 16 addresses precisely this question, ensuring that, with arbitrarily high probability, the finite particle dynamics (4) keeps close to the mean-field dynamics (8). Since the rate of this convergence is of order $N^{-1/2}$ in the number of particles N , the hardness of the problem is fully captured by the constant C_{MFA} in (31), which does not depend explicitly on the dimension d . Therefore, the mean-field approximation is, in general, not affected by the curse of dimensionality. Nevertheless, as our assumptions on the objective function \mathcal{E} do not exclude the class of NP-hard problems, it cannot be expected that CBO solves *any* problem, howsoever hard, with polynomial complexity. This is reflected by the exponential dependence of C_{MFA} on the parameter α and its possibly worst-case linear dependence on the dimension d , as we discuss in what follows. However, several numerical experiments [11, 26–28] in high dimensions confirm that in typical applications CBO performs comparably to state-of-the-art methods without the necessity of an exponentially large amount of particles. As mentioned before, characterizing α_0 in more detail is crucial in view of the mean-field approximation result, Proposition 16. We did not precisely specify α_0 in Theorem 12 since it seems challenging to provide informative bounds in all generality. In Remark 24, however, we devise an informal derivation in the case $H \equiv 1$ for objectives \mathcal{E} that are locally L -Lipschitz continuous on some ball $B_R(v^*)$ and satisfy the coercivity condition (23) globally for $\nu = 1/2$. For a parameter-dependent constant $c = c(\vartheta, \lambda, \sigma)$, we obtain

$$\alpha > \alpha_0 = \frac{-8}{c^2\eta^2\varepsilon} \log \left(\frac{c}{2\sqrt{2}} \rho_0 \left(B_{\min\{R, c^2\eta^2\varepsilon/(8L)\}}(v^*) \right) \right) \quad (33)$$

provided that the probability mass $t \mapsto \rho_t(B_{c^2\eta^2\varepsilon/(8L)}(v^*))$ is minimized at time $t = 0$. The latter assumption is motivated by numerical observations of typical successful CBO runs, where the particle density around the global minimizer tends to be minimized initially and steadily increases over time. We note that the argument of the log in (33) may induce a dependence of α_0 on the ambient dimension d , if we do not dispose of an informative initial configuration ρ_0 . For instance, if ρ_0 is measure-theoretically equivalent to the Lebesgue measure on a compact set in \mathbb{R}^d , we have $\alpha_0 \in \mathcal{O}(d \log(\varepsilon)/\varepsilon)$ as $d, 1/\varepsilon \rightarrow \infty$ by (33). If we interpreted ρ_0 as the uncertainty about the location of the global minimizer v^* , we could thus consider low-uncertainty regimes, where ρ_0 actually concentrates around v^* and α_0 may be dimension-free, or a high-uncertainty regime, where ρ_0 does not concentrate and α_0 may depend on d .

4 Proof details for Section 3.2

In this section we provide the proof details for the global convergence result of CBO in mean-field law, Theorem 12. Sections 4.1–4.3 provide auxiliary results, which might be of independent interest. In Section 4.4 we complete the proof of Theorem 12. Throughout we assume $\underline{\mathcal{E}} = 0$, which is w.l.o.g. since a constant offset to \mathcal{E} does not change the CBO dynamics.

4.1 Evolution of the mean-field limit

We now derive evolution inequalities of the energy functional $\mathcal{V}(\rho_t)$ for the cases $H \equiv 1$ and $H \neq 1$, respectively.

Lemma 17. Let $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, $H \equiv 1$, and fix $\alpha, \lambda, \sigma > 0$. Moreover, let $T > 0$ and let $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ be a weak solution to the Fokker-Planck equation (9). Then the functional $\mathcal{V}(\rho_t)$ satisfies

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\leq - (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) + \sqrt{2} (\lambda + d\sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2 \\ &\quad + \frac{d\sigma^2}{2} \|v_\alpha(\rho_t) - v^*\|_2^2. \end{aligned} \quad (34)$$

Proof. We note that the function $\phi(v) = 1/2\|v - v^*\|_2^2$ is in $\mathcal{C}_*^2(\mathbb{R}^d)$ and recall that ρ satisfies the weak solution identity (19) for all test functions in $\mathcal{C}_*^2(\mathbb{R}^d)$, see Remark 7. By applying (19) with ϕ as above, we obtain for the evolution of $\mathcal{V}(\rho_t)$

$$\frac{d}{dt} \mathcal{V}(\rho_t) = \underbrace{-\lambda \int \langle v - v^*, v - v_\alpha(\rho_t) \rangle d\rho_t(v)}_{=:T_1} + \underbrace{\frac{d\sigma^2}{2} \int \|v - v_\alpha(\rho_t)\|_2^2 d\rho_t(v)}_{=:T_2},$$

where we used $\nabla\phi(v) = v - v^*$ and $\Delta\phi(v) = d$. Expanding the right-hand side of the scalar product in the integrand of T_1 by subtracting and adding v^* yields

$$\begin{aligned} T_1 &= -\lambda \int \langle v - v^*, v - v^* \rangle d\rho_t(v) + \lambda \left\langle \int (v - v^*) d\rho_t(v), v_\alpha(\rho_t) - v^* \right\rangle \\ &\leq -2\lambda \mathcal{V}(\rho_t) + \lambda \|\mathbb{E}(\rho_t) - v^*\|_2 \|v_\alpha(\rho_t) - v^*\|_2 \end{aligned}$$

with Cauchy-Schwarz inequality being used in the last step. Similarly, again by subtracting and adding v^* , for the term T_2 we have with Cauchy-Schwarz inequality

$$T_2 \leq d\sigma^2 \left(\mathcal{V}(\rho_t) + \int \|v - v^*\|_2 d\rho_t(v) \|v_\alpha(\rho_t) - v^*\|_2 + \frac{1}{2} \|v_\alpha(\rho_t) - v^*\|_2^2 \right). \quad (35)$$

The result now follows by noting that $\|\mathbb{E}(\rho_t) - v^*\|_2 \leq \int \|v - v^*\|_2 d\rho_t(v) \leq \sqrt{2\mathcal{V}(\rho_t)}$ as a consequence of Jensen's inequality. \square

Lemma 18. Under the assumptions of Lemma 17, the functional $\mathcal{V}(\rho_t)$ satisfies

$$\frac{d}{dt} \mathcal{V}(\rho_t) \geq - (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) - \sqrt{2} (\lambda + d\sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2. \quad (36)$$

Proof. By following the lines of the proof of Lemma 17 and noticing that it holds

$$\left\langle \int (v - v^*) d\rho_t(v), v_\alpha(\rho_t) - v^* \right\rangle \geq - \|\mathbb{E}(\rho_t) - v^*\|_2 \|v_\alpha(\rho_t) - v^*\|_2$$

by Cauchy-Schwarz inequality and $\|v_\alpha(\rho_t) - v^*\|_2^2 \geq 0$, the lower bound is immediate. \square

Lemma 19. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3 and w.l.o.g. assume $\underline{\mathcal{E}} = 0$. Let $H : \mathbb{R}^d \rightarrow [0, 1]$ be such that $H(x) = 1$ whenever $x \geq 0$ and fix $\alpha, \lambda, \sigma > 0$. Moreover, let $T > 0$ and let $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ be a weak solution to the Fokker-Planck equation (9). Then, provided $\max_{t \in [0, T]} \mathcal{E}(v_\alpha(\rho_t)) \leq \mathcal{E}_\infty$, the functional $\mathcal{V}(\rho_t)$ satisfies

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\leq - (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) + \sqrt{2} (\lambda + d\sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2 \\ &\quad + \frac{\lambda}{\eta^2} (L_\mathcal{E} (1 + \|v_\alpha(\rho_t) - v^*\|_2^\gamma) \|v_\alpha(\rho_t) - v^*\|_2)^{2\nu} + \frac{d\sigma^2}{2} \|v_\alpha(\rho_t) - v^*\|_2^2. \end{aligned}$$

Proof. Let us write $H^*(v) := H(\mathcal{E}(v) - \mathcal{E}(v_\alpha(\rho_t)))$. Taking $\phi(v) = 1/2\|v - v^*\|_2^2$ as test function in (19) as in the proof of Lemma 17 yields for the evolution of $\mathcal{V}(\rho_t)$

$$\frac{d}{dt}\mathcal{V}(\rho_t) = -\lambda \underbrace{\int H^*(v)\langle v - v^*, v - v_\alpha(\rho_t) \rangle d\rho_t(v)}_{=: \tilde{T}_1} + \frac{d\sigma^2}{2} \int \|v - v_\alpha(\rho_t)\|_2^2 d\rho_t(v). \quad (37)$$

For the second term on the right-hand side, we proceed as in Equation (35). The term \tilde{T}_1 on the other hand can be rewritten as

$$\begin{aligned} \tilde{T}_1 &= -2\lambda\mathcal{V}(\rho_t) - \lambda \int H^*(v)\langle v - v^*, v^* - v_\alpha(\rho_t) \rangle d\rho_t(v) \\ &\quad + \lambda \int (1 - H^*(v)) \|v - v^*\|_2^2 d\rho_t(v). \end{aligned} \quad (38)$$

Let us now bound the latter two terms individually. For the second term in (38), noting that $0 \leq H^* \leq 1$, Cauchy-Schwarz inequality and Jensen's inequality give

$$-\lambda \int H^*(v)\langle v - v^*, v^* - v_\alpha(\rho_t) \rangle d\rho_t(v) \leq \lambda\sqrt{2\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2.$$

For the third term in (38), let us first note that $(1 - H^*(v)) \neq 0$ implies $H^*(v) \neq 1$ and thus $\mathcal{E}(v) < \mathcal{E}(v_\alpha(\rho_t))$. Furthermore, $\mathcal{E}(v_\alpha(\rho_t)) \leq \mathcal{E}_\infty$ implies $v \in B_{R_0}(v^*)$ by the second part of A2. By the first part of A2 and $0 \leq 1 - H^* \leq 1$, we therefore have

$$\begin{aligned} \lambda \int (1 - H^*(v)) \|v - v^*\|_2^2 d\rho_t(v) &\leq \lambda \int \frac{(1 - H^*(v))}{\eta^2} \mathcal{E}(v)^{2\nu} d\rho_t(v) \leq \frac{\lambda}{\eta^2} \mathcal{E}(v_\alpha(\rho_t))^{2\nu} \\ &\leq \frac{\lambda}{\eta^2} (L_{\mathcal{E}}(1 + \|v_\alpha(\rho_t) - v^*\|_2^\gamma) \|v_\alpha(\rho_t) - v^*\|_2)^{2\nu}, \end{aligned}$$

where the last step used A3. Employing the last two inequalities in (38) and inserting the result together with (35) into (37), gives the result. \square

Lemma 20. *Under the assumptions of Lemma 19, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt}\mathcal{V}(\rho_t) \geq - (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) - \sqrt{2} (\lambda + d\sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2. \quad (39)$$

Proof. Analogously to the proof of Lemma 18, by following the lines of the proof of Lemma 19 and noticing that for \tilde{T}_1 it holds

$$-\int H^*(v)\langle v - v^*, v^* - v_\alpha(\rho_t) \rangle d\rho_t(v) \geq -\|\mathbb{E}(\rho_t) - v^*\|_2 \|v_\alpha(\rho_t) - v^*\|_2$$

as well as $\int (1 - H^*(v)) \|v - v^*\|_2^2 d\rho_t(v) \geq 0$ as a consequence of $0 \leq H^* \leq 1$, the lower bound is immediate. \square

4.2 Quantitative Laplace principle

The Laplace principle (7) asserts that $-\log(\|\omega_\alpha\|_{L_1(\varrho)})/\alpha \rightarrow \mathcal{E}$ as $\alpha \rightarrow \infty$ as long as the global minimizer v^* is in the support of ϱ . However, it cannot be used to characterize the proximity of $v_\alpha(\varrho)$ to the global minimizer v^* in general. For instance, if \mathcal{E} had two minimizers with similar objective value \mathcal{E} , and half of the probability mass of ϱ concentrates around each associated location, $v_\alpha(\varrho)$ is located halfway on the line that connects the two minimizing locations. The inverse continuity property A2, by design, excludes such cases, so that we can refine the Laplace principle under A2 in the following sense.

Proposition 21. Let $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$ define $\mathcal{E}_r := \sup_{v \in B_r(v^*)} \mathcal{E}(v)$. Then, under the inverse continuity property A2 and assuming w.l.o.g. $\underline{\mathcal{E}} = 0$, for any $r \in (0, R_0]$ and $q > 0$ such that $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, we have

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \frac{(q + \mathcal{E}_r)^\nu}{\eta} + \frac{\exp(-\alpha q)}{\varrho(B_r(v^*))} \int \|v - v^*\|_2 d\varrho(v).$$

Proof. For any $a > 0$ it holds $\|\omega_\alpha\|_{L_1(\varrho)} \geq a\varrho(\{v : \exp(-\alpha\mathcal{E}(v)) \geq a\})$ due to Markov's inequality. By choosing $a = \exp(-\alpha\mathcal{E}_r)$ and noting that

$$\varrho\left(\left\{v \in \mathbb{R}^d : \exp(-\alpha\mathcal{E}(v)) \geq \exp(-\alpha\mathcal{E}_r)\right\}\right) = \varrho\left(\left\{v \in \mathbb{R}^d : \mathcal{E}(v) \leq \mathcal{E}_r\right\}\right) \geq \varrho(B_r(v^*)),$$

we get $\|\omega_\alpha\|_{L_1(\varrho)} \geq \exp(-\alpha\mathcal{E}_r)\varrho(B_r(v^*))$. Now let $\tilde{r} \geq r > 0$. Using the definition of the consensus point $v_\alpha(\varrho) = \int v\omega_\alpha(v)/\|\omega_\alpha\|_{L_1(\varrho)} d\varrho(v)$ we can decompose

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \int_{B_{\tilde{r}}(v^*)} \|v - v^*\|_2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\varrho)}} d\varrho(v) + \int_{(B_{\tilde{r}}(v^*))^c} \|v - v^*\|_2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\varrho)}} d\varrho(v).$$

The first term is bounded by \tilde{r} since $\|v - v^*\|_2 \leq \tilde{r}$ for all $v \in B_{\tilde{r}}(v^*)$. For the second term we use $\|\omega_\alpha\|_{L_1(\varrho)} \geq \exp(-\alpha\mathcal{E}_r)\varrho(B_r(v^*))$ from above to get

$$\begin{aligned} \int_{(B_{\tilde{r}}(v^*))^c} \|v - v^*\|_2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\varrho)}} d\varrho(v) &\leq \frac{1}{\exp(-\alpha\mathcal{E}_r)\varrho(B_r(v^*))} \int_{(B_{\tilde{r}}(v^*))^c} \|v - v^*\|_2 \omega_\alpha(v) d\varrho(v) \\ &\leq \frac{\exp(-\alpha(\inf_{v \in (B_{\tilde{r}}(v^*))^c} \mathcal{E}(v) - \mathcal{E}_r))}{\varrho(B_r(v^*))} \int \|v - v^*\|_2 d\varrho(v). \end{aligned}$$

Thus, for any $\tilde{r} \geq r > 0$ we obtain

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \tilde{r} + \frac{\exp(-\alpha(\inf_{v \in (B_{\tilde{r}}(v^*))^c} \mathcal{E}(v) - \mathcal{E}_r))}{\varrho(B_r(v^*))} \int \|v - v^*\|_2 d\varrho(v). \quad (40)$$

Let us now choose $\tilde{r} = (q + \mathcal{E}_r)^\nu / \eta$. This choice satisfies $\tilde{r} \leq \mathcal{E}_\infty^\nu / \eta$ by the assumption $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, and furthermore $\tilde{r} \geq r$, since A2 with $\underline{\mathcal{E}} = 0$ and $r \leq R_0$ implies

$$\tilde{r} = \frac{(q + \mathcal{E}_r)^\nu}{\eta} \geq \frac{\mathcal{E}_r^\nu}{\eta} = \frac{\left(\sup_{v \in B_r(v^*)} \mathcal{E}(v)\right)^\nu}{\eta} \geq \sup_{v \in B_r(v^*)} \|v - v^*\|_2 = r.$$

Thus, using again A2 with $\underline{\mathcal{E}} = 0$, it holds

$$\inf_{v \in (B_{\tilde{r}}(v^*))^c} \mathcal{E}(v) - \mathcal{E}_r \geq \min\{\mathcal{E}_\infty, (\eta\tilde{r})^{1/\nu}\} - \mathcal{E}_r = (\eta\tilde{r})^{1/\nu} - \mathcal{E}_r = q.$$

Inserting this and the definition of \tilde{r} into (40), we obtain the result. \square

4.3 A lower bound for the probability mass around v^*

In this section we bound the probability mass $\rho_t(B_r(v^*))$ for an arbitrary small radius $r > 0$ from below. By defining a smooth mollifier $\phi_r : \mathbb{R}^d \rightarrow [0, 1]$ with $\text{supp } \phi_r = B_r(v^*)$ according to

$$\phi_r(v) := \begin{cases} \exp\left(1 - \frac{r^2}{r^2 - \|v - v^*\|_2^2}\right), & \text{if } \|v - v^*\|_2 < r, \\ 0, & \text{else,} \end{cases} \quad (41)$$

it holds $\rho_t(B_r(v^*)) \geq \int \phi_r(v) d\rho_t(v)$. From there, the evolution of the right-hand side can be studied by using the weak solution property of ρ as in Definition 5.

Proposition 22. Let $H : \mathbb{R} \rightarrow [0, 1]$ be arbitrary, $T > 0$, $r > 0$, and fix parameters $\alpha, \lambda, \sigma > 0$. Assume $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ weakly solves the Fokker-Planck equation (9) in the sense of Definition 5 with initial condition $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ and for $t \in [0, T]$. Then, for all $t \in [0, T]$ we have

$$\rho_t(B_r(v^*)) \geq \left(\int \phi_r(v) d\rho_0(v) \right) \exp(-pt), \text{ where} \quad (42)$$

$$p := \max \left\{ \frac{2\lambda(\sqrt{cr} + B)\sqrt{c}}{(1-c)^2r} + \frac{2\sigma^2(cr^2 + B^2)(2c+d)}{(1-c)^4r^2}, \frac{4\lambda^2}{(2c-1)\sigma^2} \right\} \quad (43)$$

for any $B < \infty$ with $\sup_{t \in [0, T]} \|v_\alpha(\rho_t) - v^*\|_2 \leq B$ and for any $c \in (1/2, 1)$ satisfying

$$(2c-1)c \geq d(1-c)^2. \quad (44)$$

Remark 23. In case the reader may have wondered about the crucial role of the stochastic terms in (2) and (4), or the diffusion in the macroscopic models (8) and (9), Proposition 22 precisely explains where positive diffusion $\sigma > 0$ is actually used to ensure mass around the minimizer v^* (compare Proposition 21).

Proof of Proposition 22. By definition of the mollifier ϕ_r in (41) we have $0 \leq \phi_r(v) \leq 1$ and $\text{supp}(\phi_r) = B_r(v^*)$. This implies

$$\rho_t(B_r(v^*)) = \rho_t \left(\left\{ v \in \mathbb{R}^d : \|v - v^*\|_2 \leq r \right\} \right) \geq \int \phi_r(v) d\rho_t(v). \quad (45)$$

Our strategy is to derive a lower bound for the right-hand side of this inequality. Using the weak solution property of ρ and the fact that $\phi_r \in \mathcal{C}^\infty(\mathbb{R}^d)$, we obtain

$$\frac{d}{dt} \int \phi_r(v) d\rho_t(v) = \int (T_1(v) + T_2(v)) d\rho_t(v) \quad (46)$$

with $T_1(v) := -\lambda H^*(v) \langle v - v_\alpha(\rho_t), \nabla \phi_r(v) \rangle$ and $T_2(v) := \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \Delta \phi_r(v)$, and where we abbreviate $H^*(v) := H(\mathcal{E}(v) - \mathcal{E}(v_\alpha(\rho_t)))$ to keep the notation concise. We now aim for showing $T_1(v) + T_2(v) \geq -p\phi_r(v)$ uniformly on \mathbb{R}^d for $p > 0$ as given in (43) in the statement. Since the mollifier ϕ_r and its first and second derivatives vanish outside of $\Omega_r := \{v \in \mathbb{R}^d : \|v - v^*\|_2 < r\}$ we can restrict our attention to the open ball Ω_r . To achieve the lower bound over Ω_r , we introduce the subsets $K_1 := \{v \in \mathbb{R}^d : \|v - v^*\|_2 > \sqrt{cr}\}$ and

$$K_2 := \left\{ v \in \mathbb{R}^d : -\lambda H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle \left(r^2 - \|v - v^*\|_2^2 \right)^2 \right. \\ \left. > \tilde{c}r^2 \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2 \right\},$$

where c adheres to (44), and $\tilde{c} := 2c - 1 \in (0, 1)$. We now decompose Ω_r according to $\Omega_r = (K_1^c \cap \Omega_r) \cup (K_1 \cap K_2^c \cap \Omega_r) \cup (K_1 \cap K_2 \cap \Omega_r)$, which is illustrated in Figure 3 for different positions of $v_\alpha(\rho_t)$ and values of σ .

In the following we treat each of these three subsets separately.

Subset $K_1^c \cap \Omega_r$: We have $\|v - v^*\|_2 \leq \sqrt{cr}$ for each $v \in K_1^c$, which can be used to independently derive lower bounds for both T_1 and T_2 . Recalling the expression for ϕ_r from (41), for T_1 we get by using Cauchy-Schwarz inequality and $H^* \leq 1$

$$T_1(v) = -\lambda H^*(v) \langle v - v_\alpha(\rho_t), \nabla \phi_r(v) \rangle = -\lambda H^*(v) \left\langle v - v_\alpha(\rho_t), \frac{-2r^2(v - v^*)\phi_r(v)}{\left(r^2 - \|v - v^*\|_2^2 \right)^2} \right\rangle \\ \geq -2r^2\lambda \frac{\|v - v_\alpha(\rho_t)\|_2 \|v - v^*\|_2}{\left(r^2 - \|v - v^*\|_2^2 \right)^2} \phi_r(v) \geq -\frac{2\lambda(\sqrt{cr} + B)\sqrt{c}}{(1-c)^2r} \phi_r(v) =: -p_1\phi_r(v),$$

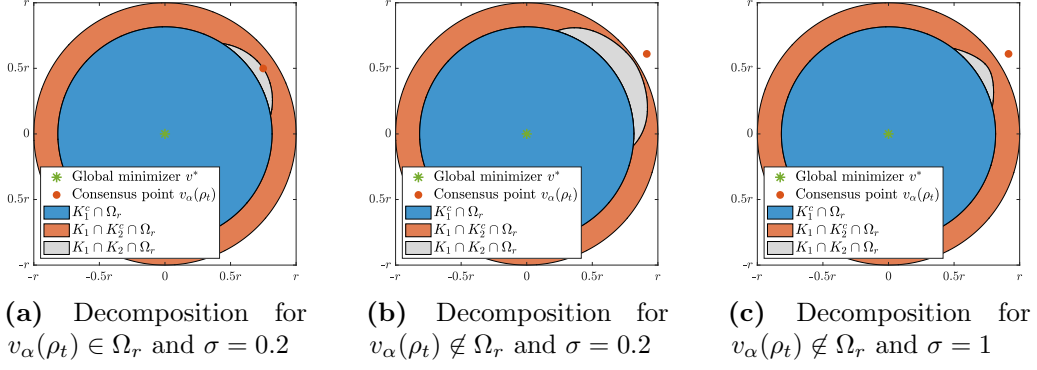


Figure 3: Visualization of the decomposition of Ω_r for different positions of $v_\alpha(\rho_t)$ and values of σ in the setting $H \equiv 1$. In the proof of Proposition 22 we limit the rate of the mass loss induced by both consensus drift and noise term for the set $K_1^c \cap \Omega_r$, which is colored blue. On the set $K_1 \cap K_2^c \cap \Omega_r$, inked orange, the noise term counterbalances any potential mass loss induced by the drift, while on the gray set $K_1 \cap K_2 \cap \Omega_r$ mass can be lost at an exponential rate $-4\lambda^2/((2c-1)\sigma^2)$.

where the last bound is due to $\|v - v_\alpha(\rho_t)\|_2 \leq \|v - v^*\|_2 + \|v^* - v_\alpha(\rho_t)\|_2 \leq \sqrt{cr} + B$. Similarly, by computing $\Delta\phi_r$ and inserting it, for T_2 we obtain

$$\begin{aligned}
T_2(v) &= \sigma^2 r^2 \|v - v_\alpha(\rho_t)\|_2^2 \frac{2 \left(2 \|v - v^*\|_2^2 - r^2 \right) \|v - v^*\|_2^2 - d \left(r^2 - \|v - v^*\|_2^2 \right)^2}{\left(r^2 - \|v - v^*\|_2^2 \right)^4} \phi_r(v) \\
&\geq -\frac{2\sigma^2 (cr^2 + B^2)(2c + d)}{(1-c)^4 r^2} \phi_r(v) =: -p_2 \phi_r(v),
\end{aligned}$$

where we used $\|v - v_\alpha(\rho_t)\|_2^2 \leq 2(\|v - v^*\|_2^2 + \|v^* - v_\alpha(\rho_t)\|_2^2) \leq 2(cr^2 + B^2)$.

Subset $K_1 \cap K_2^c \cap \Omega_r$: By the definition of K_1 and K_2 we have $\|v - v^*\|_2 > \sqrt{cr}$ and

$$-\lambda H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle \left(r^2 - \|v - v^*\|_2^2 \right)^2 \leq \tilde{c} r^2 \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2. \quad (47)$$

Our goal now is to show $T_1(v) + T_2(v) \geq 0$ for all v in this subset. We first compute

$$\begin{aligned}
\frac{T_1(v) + T_2(v)}{2r^2 \phi_r(v)} &= \lambda H^*(v) \frac{\langle v - v_\alpha(\rho_t), v - v^* \rangle \left(r^2 - \|v - v^*\|_2^2 \right)^2}{\left(r^2 - \|v - v^*\|_2^2 \right)^4} \\
&\quad + \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \frac{2 \left(2 \|v - v^*\|_2^2 - r^2 \right) \|v - v^*\|_2^2 - d \left(r^2 - \|v - v^*\|_2^2 \right)^2}{\left(r^2 - \|v - v^*\|_2^2 \right)^4}.
\end{aligned}$$

Therefore we have $T_1(v) + T_2(v) \geq 0$ whenever we can show

$$\begin{aligned}
&\left(-\lambda H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle + \frac{d\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \right) \left(r^2 - \|v - v^*\|_2^2 \right)^2 \\
&\leq \sigma^2 \|v - v_\alpha(\rho_t)\|_2^2 \left(2 \|v - v^*\|_2^2 - r^2 \right) \|v - v^*\|_2^2.
\end{aligned} \quad (48)$$

Now note that the first summand on the left-hand side in (48) can be upper bounded by means

of Condition (47) and by using the relation $\tilde{c} = 2c - 1$. More precisely,

$$\begin{aligned} -\lambda H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle \left(r^2 - \|v - v^*\|_2^2 \right)^2 &\leq \tilde{c} r^2 \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2 \\ &= (2c - 1) r^2 \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2 \leq \left(2 \|v - v^*\|_2^2 - r^2 \right) \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2, \end{aligned}$$

where the last inequality follows since $v \in K_1$. For the second term on the left-hand side in (48) we can use $d(1 - c)^2 \leq (2c - 1)c$ as per (44), to get

$$\begin{aligned} \frac{d\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \left(r^2 - \|v - v^*\|_2^2 \right)^2 &\leq \frac{d\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 (1 - c)^2 r^4 \\ &\leq \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 (2c - 1) r^2 c r^2 \leq \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \left(2 \|v - v^*\|_2^2 - r^2 \right) \|v - v^*\|_2^2. \end{aligned}$$

Hence, (48) holds and we have $T_1(v) + T_2(v) \geq 0$ uniformly on this subset.

Subset $K_1 \cap K_2 \cap \Omega_r$: On this subset, we have $\|v - v^*\|_2 > \sqrt{cr}$ and

$$-\lambda H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle \left(r^2 - \|v - v^*\|_2^2 \right)^2 > \tilde{c} r^2 \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2. \quad (49)$$

We first note that $T_1(v) = 0$ whenever $\sigma^2 \|v - v_\alpha(\rho_t)\|_2^2 = 0$, provided that $\sigma > 0$, so nothing needs to be done for the point $v = v_\alpha(\rho_t)$. On the other hand, if $\sigma^2 \|v - v_\alpha(\rho_t)\|_2^2 > 0$, we can use $H^* \leq 1$, two applications of Cauchy-Schwarz inequalities, and Condition (49) to get

$$\begin{aligned} \frac{H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle}{\left(r^2 - \|v - v^*\|_2^2 \right)^2} &\geq \frac{-\|v - v_\alpha(\rho_t)\|_2 \|v - v^*\|_2}{\left(r^2 - \|v - v^*\|_2^2 \right)^2} \\ &> \frac{2\lambda H^*(v) \langle v - v_\alpha(\rho_t), v - v^* \rangle}{\tilde{c} r^2 \sigma^2 \|v - v_\alpha(\rho_t)\|_2 \|v - v^*\|_2} \geq -\frac{2\lambda}{\tilde{c} r^2 \sigma^2}. \end{aligned}$$

Using this, T_1 can be bounded from below by

$$T_1(v) = 2\lambda r^2 H^*(v) \left\langle v - v_\alpha(\rho_t), \frac{v - v^*}{\left(r^2 - \|v - v^*\|_2^2 \right)^2} \phi_r(v) \right\rangle \geq -\frac{4\lambda^2}{\tilde{c}\sigma^2} \phi_r(v) =: -p_3 \phi_r(v),$$

where we made use of the relation $\tilde{c} = 2c - 1$ in the last step. For T_2 , we note that the nonnegativity of $\sigma^2 \|v - v_\alpha(\rho_t)\|_2$ implies $T_2(v) \geq 0$, whenever

$$2 \left(2 \|v - v^*\|_2^2 - r^2 \right) \|v - v^*\|_2^2 \geq d \left(r^2 - \|v - v^*\|_2^2 \right)^2.$$

This is satisfied for all v with $\|v - v^*\|_2 \geq \sqrt{cr}$, provided c satisfies $2(2c - 1)c \geq (1 - c)^2 d$ as implied by (44).

Concluding the proof: Using the evolution of ϕ_r as in (46), we now get

$$\begin{aligned} \frac{d}{dt} \int \phi_r(v) d\rho_t(v) &= \int_{K_1 \cap K_2^c \cap \Omega_r} \underbrace{(T_1(v) + T_2(v))}_{\geq 0} d\rho_t(v) \\ &\quad + \int_{K_1 \cap K_2 \cap \Omega_r} \underbrace{(T_1(v) + T_2(v))}_{\geq -p_3 \phi_r(v)} d\rho_t(v) + \int_{K_1^c \cap \Omega_r} \underbrace{(T_1(v) + T_2(v))}_{\geq -(p_1 + p_2) \phi_r(v)} d\rho_t(v) \\ &\geq -\max\{p_1 + p_2, p_3\} \int \phi_r(v) d\rho_t(v) = -p \int \phi_r(v) d\rho_t(v) \end{aligned}$$

An application of Grönwall's inequality gives $\int \phi_r(v) d\rho_t(v) \geq \int \phi_r(v) d\rho_0(v) \exp(-pt)$, which concludes the proof after recalling (45). \square

4.4 Proof of Theorem 12

We now have all necessary tools at hand to present a detailed proof of the global convergence result in mean-field law. We separately prove the cases of an inactive and active cutoff function, i.e., $H \equiv 1$ and $H \not\equiv 1$, respectively.

Proof of Theorem 12 when $H \equiv 1$. W.l.o.g. we may assume $\underline{\mathcal{E}} = 0$. Let us first choose the parameter α such that

$$\alpha > \alpha_0 := \frac{1}{q_\varepsilon} \left(\log \left(\frac{4\sqrt{2\mathcal{V}(\rho_0)}}{c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon}} \right) + \frac{p}{(1-\vartheta)(2\lambda - d\sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) - \log \rho_0(B_{r_\varepsilon/2}(v^*)) \right), \quad (50)$$

where we introduce the definitions

$$c(\vartheta, \lambda, \sigma) := \min \left\{ \frac{\vartheta}{2} \frac{(2\lambda - d\sigma^2)}{\sqrt{2}(\lambda + d\sigma^2)}, \sqrt{\frac{\vartheta(2\lambda - d\sigma^2)}{d\sigma^2}} \right\} \quad (51)$$

as well as

$$q_\varepsilon := \frac{1}{2} \min \left\{ \left(\eta \frac{c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon}}{2} \right)^{1/\nu}, \mathcal{E}_\infty \right\} \quad \text{and} \quad r_\varepsilon := \max_{s \in [0, R_0]} \left\{ \max_{v \in B_s(v^*)} \mathcal{E}(v) \leq q_\varepsilon \right\}.$$

Moreover, p is as defined in (43) in Proposition 22 with $B = c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_0)}$ and with $r = r_\varepsilon$. We remark that, by construction, $q_\varepsilon > 0$ and $r_\varepsilon \leq R_0$. Furthermore, recalling the notation $\mathcal{E}_r = \sup_{v \in B_r(v^*)} \mathcal{E}(v)$ from Proposition 21, we have $q_\varepsilon + \mathcal{E}_{r_\varepsilon} \leq 2q_\varepsilon \leq \mathcal{E}_\infty$ as a consequence of the definition of r_ε . Since $q_\varepsilon > 0$, the continuity of \mathcal{E} ensures that there exists $s_{q_\varepsilon} > 0$ such that $\mathcal{E}(v) \leq q_\varepsilon$ for all $v \in B_{s_{q_\varepsilon}}(v^*)$, thus yielding also $r_\varepsilon > 0$.

Let us now define the time horizon $T_\alpha \geq 0$, which may depend on α , by

$$T_\alpha := \sup \left\{ t \geq 0 : \mathcal{V}(\rho_t) > \varepsilon \text{ and } \|v_\alpha(\rho_t) - v^*\|_2 < C(t) \text{ for all } t' \in [0, t] \right\} \quad (52)$$

with $C(t) := c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_t)}$. Notice for later use that $C(0) = B$.

Our aim now is to show $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in [\frac{1-\vartheta}{(1+\vartheta/2)} T^*, T^*]$ and that we have at least exponential decay of $\mathcal{V}(\rho_t)$ until time T_α , i.e., until accuracy ε is reached.

First, however, we ensure that $T_\alpha > 0$. With the mapping $t \mapsto \mathcal{V}(\rho_t)$ being continuous as a consequence of the regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ established in Theorem 6 and $t \mapsto \|v_\alpha(\rho_t) - v^*\|_2$ being continuous due to [10, Lemma 3.2] and $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$, $T_\alpha > 0$ follows from the definition, since $\mathcal{V}(\rho_0) > \varepsilon$ and $\|v_\alpha(\rho_0) - v^*\|_2 < C(0)$. While the former is immediate by assumption, applying Proposition 21 with q_ε and r_ε gives the latter since

$$\begin{aligned} \|v_\alpha(\rho_0) - v^*\|_2 &\leq \frac{(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\exp(-\alpha q_\varepsilon)}{\rho_0(B_{r_\varepsilon}(v^*))} \int \|v - v^*\|_2 d\rho_0(v) \\ &\leq \frac{c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon}}{2} + \frac{\exp(-\alpha q_\varepsilon)}{\rho_0(B_{r_\varepsilon}(v^*))} \sqrt{2\mathcal{V}(\rho_0)} \\ &\leq c(\vartheta, \lambda, \sigma) \sqrt{\varepsilon} < c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_0)} = C(0), \end{aligned}$$

where the first inequality in the last line holds by the choice of α in (50).

Next, we show that the functional $\mathcal{V}(\rho_t)$ decays essentially exponentially fast in time. More precisely, we prove that, up to time T_α , $\mathcal{V}(\rho_t)$ decays

- (i) at least exponentially fast (with rate $(1-\vartheta)(2\lambda - d\sigma^2)$), and

(ii) at most exponentially fast (with rate $(1 + \vartheta/2)(2\lambda - d\sigma^2)$).

To obtain (i), recall that Lemma 17 provides an upper bound on $\frac{d}{dt}\mathcal{V}(\rho_t)$ given by

$$\begin{aligned} \frac{d}{dt}\mathcal{V}(\rho_t) &\leq - (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) + \sqrt{2} (\lambda + d\sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2 \\ &\quad + \frac{d\sigma^2}{2} \|v_\alpha(\rho_t) - v^*\|_2^2. \end{aligned} \quad (53)$$

Combining this with the definition of T_α in (52) we have by construction

$$\frac{d}{dt}\mathcal{V}(\rho_t) \leq -(1 - \vartheta) (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) \quad \text{for all } t \in (0, T_\alpha).$$

Analogously, for (ii), by Lemma 18, we obtain a lower bound on $\frac{d}{dt}\mathcal{V}(\rho_t)$ of the form

$$\begin{aligned} \frac{d}{dt}\mathcal{V}(\rho_t) &\geq - (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) - \sqrt{2} (\lambda + d\sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2 \\ &\geq -(1 + \vartheta/2) (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) \quad \text{for all } t \in (0, T_\alpha), \end{aligned}$$

where the second inequality again exploits the definition of T_α . Grönwall's inequality now implies for all $t \in [0, T_\alpha]$ the upper and lower bound

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0) \exp(- (1 - \vartheta) (2\lambda - d\sigma^2) t), \quad (54)$$

$$\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0) \exp(- (1 + \vartheta/2) (2\lambda - d\sigma^2) t), \quad (55)$$

thereby proving (i) and (ii). We further note that the definition of T_α in (52) together with the definition of $C(t)$ and (54) permits to control

$$\max_{t \in [0, T_\alpha]} \|v_\alpha(\rho_t) - v^*\|_2 \leq \max_{t \in [0, T_\alpha]} C(t) \leq C(0). \quad (56)$$

To conclude, it remains to prove that $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in [\frac{1-\vartheta}{(1+\vartheta/2)} T^*, T^*]$. For this we distinguish the following three cases.

Case $T_\alpha \geq T^*$: We can use the definition of T^* in (26) and the time-evolution bound of $\mathcal{V}(\rho_t)$ in (54) to conclude that $\mathcal{V}(\rho_{T^*}) \leq \varepsilon$. Hence, by definition of T_α in (52) together with the continuity of $\mathcal{V}(\rho_t)$, we find $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha = T^*$.

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) \leq \varepsilon$: By continuity of $\mathcal{V}(\rho_t)$, it holds for T_α as defined in (52), $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$. Thus, $\varepsilon = \mathcal{V}(\rho_{T_\alpha}) \geq \mathcal{V}(\rho_0) \exp(- (1 + \vartheta/2) (2\lambda - d\sigma^2) T_\alpha)$ by (55), which can be reordered as

$$\frac{1 - \vartheta}{(1 + \vartheta/2)} T^* = \frac{1}{(1 + \vartheta/2) (2\lambda - d\sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) \leq T_\alpha < T^*.$$

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$: We shall show that this case can never occur by verifying that $\|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 < C(T_\alpha)$ due to the choice of α in (50). In fact, fulfilling simultaneously both $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$ and $\|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 < C(T_\alpha)$ would contradict the definition of T_α in (52) itself. To this end, by applying again Proposition 21 with q_ε and r_ε , and recalling that $\varepsilon < \mathcal{V}(\rho_{T_\alpha})$, we get

$$\begin{aligned} \|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 &\leq \frac{(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\exp(-\alpha q_\varepsilon)}{\rho_{T_\alpha}(B_{r_\varepsilon}(v^*))} \int \|v - v^*\|_2 d\rho_{T_\alpha}(v) \\ &< \frac{c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{\exp(-\alpha q_\varepsilon)}{\rho_{T_\alpha}(B_{r_\varepsilon}(v^*))} \sqrt{2\mathcal{V}(\rho_{T_\alpha})}. \end{aligned} \quad (57)$$

Since, thanks to (56), we have the bound $\max_{t \in [0, T_\alpha]} \|v_\alpha(\rho_t) - v^*\|_2 \leq B$ for $B = C(0)$, which is in particular independent of α , Proposition 22 guarantees that there exists a $p > 0$ not depending on α (but depending on B and r_ε) with

$$\rho_{T_\alpha}(B_{r_\varepsilon}(v^*)) \geq \left(\int \phi_{r_\varepsilon}(v) d\rho_0(v) \right) \exp(-pT_\alpha) \geq \frac{1}{2} \rho_0(B_{r_\varepsilon/2}(v^*)) \exp(-pT^*) > 0,$$

where we used $v^* \in \text{supp}(\rho_0)$ for bounding the initial mass ρ_0 and the fact that ϕ_r (as defined in Equation (41)) is bounded from below on $B_{r/2}(v^*)$ by $1/2$. With this we can continue the chain of inequalities in (57) to obtain

$$\begin{aligned} \|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 &< \frac{c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2 \exp(-\alpha q_\varepsilon)}{\rho_0(B_{r_\varepsilon/2}(v^*)) \exp(-pT^*)} \sqrt{2\mathcal{V}(\rho_{T_\alpha})} \\ &\leq c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_{T_\alpha})} = C(T_\alpha), \end{aligned} \quad (58)$$

where the first inequality in the last line holds by the choice of α in (50). This establishes the desired contradiction, again as consequence of the continuity of the mappings $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \|v_\alpha(\rho_t) - v^*\|_2$. \square

Proof of Theorem 12 when $H \neq 1$. The proof follows the lines of the one for the inactive cutoff $H \equiv 1$, but requires some modifications since Lemmas 17 and 18 need to be replaced by Lemmas 19 and 20, to derive bounds for the evolution of $\mathcal{V}(\rho_t)$.

As in the proof for the case $H \equiv 1$ we first choose the parameter α such that

$$\begin{aligned} \alpha > \alpha_0 := \frac{1}{q_\varepsilon} \left(\log \left(\frac{4\sqrt{2\mathcal{V}(\rho_0)}}{C_\varepsilon} \right) + \frac{p}{(1-\vartheta)(2\lambda - d\sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) \right. \\ \left. - \log \rho_0(B_{r_\varepsilon/2}(v^*)) \right), \end{aligned} \quad (59)$$

where C_ε is obtained when replacing with ε each $\mathcal{V}(\rho_t)$ in $C(t)$ defined as

$$\begin{aligned} C(t) := \min \left\{ \frac{\mathcal{E}_\infty}{2L_\mathcal{E}}, \left(\frac{\mathcal{E}_\infty}{2L_\mathcal{E}} \right)^{1/(1+\gamma)}, \frac{\vartheta}{4} \frac{(2\lambda - d\sigma^2)}{\sqrt{2}(\lambda + d\sigma^2)} \sqrt{\mathcal{V}(\rho_t)}, \sqrt{\frac{\vartheta}{2} \frac{(2\lambda - d\sigma^2)}{d\sigma^2} \mathcal{V}(\rho_t)}, \right. \\ \left. \left(\frac{\vartheta}{4} \frac{\eta^2}{L_\mathcal{E}^{2\nu}} \frac{(2\lambda - d\sigma^2)}{\lambda} \mathcal{V}(\rho_t) \right)^{1/(2\nu)}, \left(\frac{\vartheta}{4} \frac{\eta^2}{L_\mathcal{E}^{2\nu}} \frac{(2\lambda - d\sigma^2)}{\lambda} \mathcal{V}(\rho_t) \right)^{1/(2\nu(1+\gamma))} \right\}. \end{aligned} \quad (60)$$

Moreover, r_ε is as defined before, p as in (43) with $B = C(0)$ and $r = r_\varepsilon$, and

$$q_\varepsilon := \frac{1}{2} \min \left\{ \left(\eta \frac{C_\varepsilon}{2} \right)^{1/\nu}, \mathcal{E}_\infty \right\}.$$

Let us now define again a time horizon T_α according to (52), however with the modified definition of $C(t)$ from (60). It is straightforward to check that $T_\alpha > 0$ by choice of α in (59). Our aim is again to show $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in \left[\frac{1-\vartheta}{(1+\vartheta/2)} T^*, T^* \right]$ and that we have at least exponential decay of $\mathcal{V}(\rho_t)$ until T_α .

Since due to Assumption A3 and with the definition of $C(t)$ in (60) it holds

$$\max_{t \in [0, T_\alpha]} \mathcal{E}(v_\alpha(\rho_t)) \leq \max_{t \in [0, T_\alpha]} L_\mathcal{E} (1 + \|v_\alpha(\rho_t) - v^*\|_2^2) \|v_\alpha(\rho_t) - v^*\|_2 \leq \mathcal{E}_\infty, \quad (61)$$

Lemmas 19 and 20 provide an upper and a lower bound for the time derivative of $\mathcal{V}(\rho_t)$, which, when being combined with the definitions of T_α and $C(t)$ in (60), yield

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -(1-\vartheta) (2\lambda - d\sigma^2) \mathcal{V}(\rho_t) \quad \text{and} \quad \frac{d}{dt} \mathcal{V}(\rho_t) \geq -(1+\vartheta/2) (2\lambda - d\sigma^2) \mathcal{V}(\rho_t)$$

for all $t \in (0, T_\alpha)$ as before. We can thus follow the lines of the proof for the case $H \equiv 1$, since also here $C(t)$ is bounded. In particular, the choice of α in (59) allows to derive the contradiction $\|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 < C(T_\alpha)$ by employing Propositions 21 and 22. \square

Remark 24 (Informal lower bound for α_0). As mentioned in Section 3.3, insightful lower bounds on the required α_0 in Theorem 12 may be interesting in view of better understanding the convergence of the microscopic system (4) to the mean-field limit (8). Let us therefore informally derive in what follows an instructive lower bound on the required α_0 under the assumption that \mathcal{E} satisfies Condition A2 globally with $\nu = 1/2$ and that \mathcal{E} is locally L -Lipschitz continuous around v^* , i.e., in some ball $B_R(v^*)$. We restrict ourselves to the case of an inactive cutoff function $H \equiv 1$.

Recalling (53) in the proof of Theorem 12, α should be large enough to ensure

$$\|v_\alpha(\rho_t) - v^*\|_2 \leq c(\vartheta, \lambda, \sigma) \sqrt{\mathcal{V}(\rho_t)} \quad \text{for all } t \in [0, T], \quad (62)$$

where T is the time satisfying $\mathcal{V}(\rho_T) = \varepsilon$. To achieve this, we recall that for $\varrho \in \mathcal{P}(\mathbb{R}^d)$ the quantitative Laplace principle in Proposition 21 with choices $q_\varepsilon := c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon / 8$ and $r_\varepsilon := \min\{R, q_\varepsilon / L\}$ for q and r , respectively, yields

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \frac{\sqrt{2q_\varepsilon}}{\eta} + \frac{\exp(-\alpha q_\varepsilon)}{\varrho(B_{r_\varepsilon}(v^*))} \int \|v - v^*\|_2 d\varrho(v)$$

provided that A2 holds globally with $\nu = 1/2$ and that \mathcal{E} is L -Lipschitz continuous on some ball $B_R(v^*)$. It remains to choose $\alpha > \alpha_0$, where

$$\alpha_0 := \sup_{t \in [0, T]} \frac{-8}{c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon} \log \left(\frac{c(\vartheta, \lambda, \sigma)}{2\sqrt{2}} \rho_t \left(B_{\min\{R, c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon / (8L)\}}(v^*) \right) \right), \quad (63)$$

suggesting that α_0 is strongly related to the time-evolution of the probability mass of ρ_t around v^* . Recalling Proposition 22, this mass adheres to the lower bound

$$\rho_t(B_r(v^*)) \geq \rho_0(B_{r/2}(v^*)) \exp(-pt)/2 \quad \text{for some } p > 0 \text{ and any } r > 0.$$

However, this result is pessimistic due to its worst-case nature, and inserting it into (63) with the corresponding p as in (43) leads to overly stringent requirements on α_0 , which are reflected by the respective second summands in (50) and (59). Rather, a successful application of the CBO method entails that the probability mass around the global minimizer increases over time, so that $t \mapsto \rho_t(B_r(v^*))$ is typically minimized at $t = 0$. In such case, the lower bound (63) becomes

$$\alpha_0 = \frac{-8}{c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon} \log \left(\frac{c(\vartheta, \lambda, \sigma)}{2\sqrt{2}} \rho_0 \left(B_{\min\{R, c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon / (8L)\}}(v^*) \right) \right). \quad (64)$$

5 Proof details for Section 3.3

In this section we provide the proof details for the result about the mean-field approximation of CBO, Proposition 16. After giving the proof of the auxiliary Lemma 15, which ensures that the dynamics is to some extent bounded, we prove Proposition 16.

Proof of Lemma 15. By combining the ideas of [10, Lemma 3.4] with a Doob-like inequality, we derive a bound for $\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max \{ \|V_t^i\|_2^4 + \|\bar{V}_t^i\|_2^4 \}$, which ensures that $\hat{\rho}_t^N, \bar{\rho}_t^N \in \mathcal{P}_4(\mathbb{R}^d)$ with high probability. Here, $\bar{\rho}^N$ denotes the empirical measure associated with the processes $(\bar{V}^i)_{i=1, \dots, N}$.

Employing standard inequalities shows

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \|V_t^i\|_2^4 &\lesssim \mathbb{E} \|V_0^i\|_2^4 + \lambda^4 \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t (V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)) d\tau \right\|_2^4 \\ &\quad + \sigma^4 \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t \|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\|_2 dB_\tau^i \right\|_2^4, \end{aligned} \quad (65)$$

where we note that the expression $\int_0^t \|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\|_2 dB_\tau^i$ appearing in the third term of the right-hand side is a martingale, which is a consequence of [51, Corollary 3.2.6] combined with the regularity established in [10, Lemma 3.4]. This allows to apply the Burkholder-Davis-Gundy inequality [56, Chapter IV, Theorem 4.1], which yields

$$\mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t \|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\|_2 dB_\tau^i \right\|_2^4 \lesssim \mathbb{E} \left(\int_0^T \|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\|_2^2 d\tau \right)^2.$$

Let us stress that the constant appearing in the latter estimate depends on the dimension d . Further bounding this as well as the second term of the right-hand side in (65) by means of Jensen's inequality and utilizing [10, Lemma 3.3] yields

$$\mathbb{E} \sup_{t \in [0, T]} \|V_t^i\|_2^4 \leq C \left(1 + \mathbb{E} \|V_0^i\|_2^4 + \mathbb{E} \int_0^T \|V_\tau^i\|_2^4 + \int \|v\|_2^4 d\widehat{\rho}_\tau^N(v) d\tau \right) \quad (66)$$

with a constant $C = C(\lambda, \sigma, d, T, b_1, b_2)$. Averaging (66) over i allows to bound

$$\mathbb{E} \sup_{t \in [0, T]} \int \|v\|_2^4 d\widehat{\rho}_t^N(v) \leq C \left(1 + \mathbb{E} \int \|v\|_2^4 d\widehat{\rho}_0^N(v) + 2 \int_0^T \mathbb{E} \sup_{\hat{\tau} \in [0, \tau]} \int \|v\|_2^4 d\widehat{\rho}_\tau^N(v) d\tau \right),$$

which, after applying Grönwall's inequality, ensures that the left-hand side is bounded independently of N by a constant $K = K(\lambda, \sigma, d, T, b_1, b_2)$. With analogous arguments,

$$\mathbb{E} \sup_{t \in [0, T]} \int \|v\|_2^4 d\bar{\rho}_t^N(v) \leq K.$$

Equation (30) follows now from Markov's inequality. \square

Proof of Proposition 16. By exploiting the boundedness thanks to Lemma 15 through a cutoff technique, we can follow the steps taken in [25, Theorem 3.1].

Let us define the cutoff function

$$I_M(t) = \begin{cases} 1, & \text{if } \frac{1}{N} \sum_{i=1}^N \max \left\{ \|V_\tau^i\|_2^4, \|\bar{V}_\tau^i\|_2^4 \right\} \leq M \text{ for all } \tau \in [0, t], \\ 0, & \text{else,} \end{cases}$$

which is adapted to the natural filtration and has the property $I_M(t) = I_M(t)I_M(\tau)$ for all $\tau \in [0, t]$. With Jensen's inequality and Itô isometry this allows to derive

$$\mathbb{E} \|V_t^i - \bar{V}_t^i\|_2^2 I_M(t) \lesssim c \int_0^t \mathbb{E} \left(\|V_\tau^i - \bar{V}_\tau^i\|_2^2 + \|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 \right) I_M(\tau) d\tau \quad (67)$$

for $c = (\lambda^2 T + \sigma^2)$. Here we directly used that the processes V_t^i and \bar{V}_t^i share the initial data as well as the Brownian motion paths. In what follows, let us denote by $\bar{\rho}_\tau^N$ the empirical measure of the processes \bar{V}_τ^i . Then, by using the same arguments as in the proofs of [10, Lemma 3.2] and

[25, Lemma 3.1] with the care of taking into consideration the multiplication with the random variable $I_M(\tau)$, we obtain

$$\begin{aligned} \mathbb{E} \|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau) &\lesssim \mathbb{E} \|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\|_2^2 I_M(\tau) + \mathbb{E} \|v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau) \\ &\leq C \left(\max_{i=1,\dots,N} \mathbb{E} \|V_\tau^i - \bar{V}_\tau^i\|_2^2 I_M(\tau) + N^{-1} \right) \end{aligned}$$

for a constant $C = C(\alpha, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. After plugging the latter into (67) and taking the maximum over i , the quantitative mean-field approximation result (31) follows from an application of Grönwall’s inequality after recalling the definition of the conditional expectation and noting that $\mathbb{1}_{\Omega_M} \leq I_M(t)$ pointwise and for all $t \in [0, T]$. \square

6 Conclusions

In this paper we establish the convergence of consensus-based optimization (CBO) methods to the global minimizer. The proof technique is based on the novel insight that the dynamics of individual agents follow, on average over all realizations of Brownian motion paths, the gradient flow dynamics associated with the map $v \mapsto \|v - v^*\|_2^2$, where v^* is the global minimizer of the objective \mathcal{E} . This implies that CBO methods are barely influenced by the local energy landscape of \mathcal{E} , suggesting a high degree of robustness and versatility of the method. As opposed to restrictive concentration conditions on the initial agent configuration ρ_0 in the analyses in [10, 26, 31, 32], our result holds under mild assumptions about the initial distribution ρ_0 . Furthermore, we merely require local Lipschitz continuity and a certain tractability condition about the objective \mathcal{E} , relaxing the regularity requirement $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ together with further assumptions from prior works. In order to demonstrate the relevance of the result of convergence in mean-field law for establishing a complete convergence proof of the original numerical scheme (2), we prove a probabilistic quantitative result about the mean-field approximation, which connects the finite particle regime with the mean-field limit. With this we close the gap regarding the mean-field approximation of CBO and provide the first, and so far unique, holistic convergence proof of CBO on the plane.

We believe that the proposed analysis strategy can be adopted to other recently developed adaptations of the CBO algorithm, such as CBO methods tailored to manifold optimization problems [25, 26], polarized CBO adjusted to identify multiple minimizers simultaneously [9], as well as related metaheuristics including, for instance, Particle Swarm Optimization [30, 38, 42], which can be regarded as a second-order variant of CBO with inertia [20, 30]. For CBO with anisotropic Brownian motions, which are especially relevant in high-dimensional optimization problems [11], for CBO with memory effects and gradient information, which can be beneficial in signal processing and machine learning applications [13, 57], for CBO reconfigured for multi-objective optimization, as well as for constrained CBO, this has already been done in [28], [57], [7], and [8], respectively.

Acknowledgments

The authors would like to profusely thank Hui Huang for many fruitful and stimulating discussions about the topic.

This work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content. MF further acknowledges the support of the DFG Project “Identification of Energies from Observations of Evolutions” and the DFG SPP 1962 “Non-smooth and Complementarity-Based Distributed Parameter Systems: Simulation and Hierarchical Optimization”. TK acknowledges the support of the Technical University of Munich for hosting

him while conducting the work on this manuscript. KR acknowledges the financial support from the Technical University of Munich – Institute for Ethics in Artificial Intelligence (IEAI).

References

- [1] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines. A stochastic approach to combinatorial optimization and neural computing*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Ltd., Chichester, 1989.
- [2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
- [3] M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. Optim.*, 10(4):1116–1135, 2000.
- [4] T. Bäck, D. B. Fogel, and Z. Michalewicz, editors. *Handbook of evolutionary computation*. Institute of Physics Publishing, Bristol; Oxford University Press, New York, 1997.
- [5] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*, 35(3):268–308, 2003.
- [6] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.*, 165(2, Ser. A):471–507, 2017.
- [7] G. Borghi, M. Herty, and L. Pareschi. An adaptive consensus based method for multi-objective optimization with uniform Pareto front approximation. *Appl. Math. Optim.*, 88(2):Paper No. 58, 43, 2023.
- [8] G. Borghi, M. Herty, and L. Pareschi. Constrained consensus-based optimization. *SIAM J. Optim.*, 33(1):211–236, 2023.
- [9] L. Bungert, P. Wacker, and T. Roith. Polarized consensus-based dynamics for optimization and sampling. *arXiv preprint arXiv:2211.05238*, 2022.
- [10] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28(6):1037–1066, 2018.
- [11] J. A. Carrillo, S. Jin, L. Li, and Y. Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.*, 27(suppl.):Paper No. S5, 22, 2021.
- [12] J. A. Carrillo, C. Totzeck, and U. Vaes. Consensus-based optimization and ensemble kalman inversion for global optimization problems with constraints. In *Modeling and Simulation for Collective Dynamics*, pages 195–230. World Scientific, 2023.
- [13] J. A. Carrillo, N. G. Trillos, S. Li, and Y. Zhu. FedCBO: Reaching group consensus in clustered federated learning through consensus-based optimization. *arXiv preprint arXiv:2305.02894*, 2023.
- [14] L.-P. Chaintron and A. Diez. Propagation of chaos: a review of models, methods and applications. I. Models and methods. *Kinet. Relat. Models*, 15(6):895–1015, 2022.
- [15] L.-P. Chaintron and A. Diez. Propagation of chaos: a review of models, methods and applications. II. Applications. *Kinet. Relat. Models*, 15(6):1017–1173, 2022.

- [16] J. Chen, S. Jin, and L. Lyu. A consensus-based global optimization method with adaptive momentum estimation. *Commun. Comput. Phys.*, 31(4):1296–1316, 2022.
- [17] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. Diffusion for global optimization in \mathbb{R}^n . *SIAM J. Control Optim.*, 25(3):737–753, 1987.
- [18] L. Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Trans. Mach. Learn. Res.*, 2022, 2022.
- [19] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [20] C. Cipriani, H. Huang, and J. Qiu. Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Math. Anal.*, 54(3):3091–3121, 2022.
- [21] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.
- [22] J. Dong and X. T. Tong. Replica exchange for non-convex optimization. *J. Mach. Learn. Res.*, 22:Paper No. 173, 59, 2021.
- [23] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.
- [24] D. B. Fogel. *Evolutionary computation. Toward a new philosophy of machine intelligence*. IEEE Press, Piscataway, NJ, second edition, 2000.
- [25] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.*, 30(14):2725–2751, 2020.
- [26] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.*, 22:Paper No. 237, 55, 2021.
- [27] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Anisotropic diffusion in consensus-based optimization on the sphere. *SIAM J. Optim.*, 32(3):1984–2012, 2022.
- [28] M. Fornasier, T. Klock, and K. Riedl. Convergence of anisotropic consensus-based optimization in mean-field law. In J. L. J. Laredo, J. I. Hidalgo, and K. O. Babaagba, editors, *Applications of Evolutionary Computation - 25th European Conference, EvoApplications 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings*, volume 13224 of *Lecture Notes in Computer Science*, pages 738–754. Springer, 2022.
- [29] N. J. Gerber, F. Hoffmann, and U. Vaes. Mean-field limits for consensus-based optimization and sampling. *arXiv preprint arXiv:2312.07373*, 2023.
- [30] S. Grassi and L. Pareschi. From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Math. Models Methods Appl. Sci.*, 31(8):1625–1657, 2021.
- [31] S.-Y. Ha, S. Jin, and D. Kim. Convergence of a first-order consensus-based global optimization algorithm. *Math. Models Methods Appl. Sci.*, 30(12):2417–2444, 2020.
- [32] S.-Y. Ha, S. Jin, and D. Kim. Convergence and error estimates for time-discrete consensus-based optimization algorithms. *Numer. Math.*, 147(2):255–282, 2021.

- [33] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [34] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001.
- [35] J. H. Holland. *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor, Mich., 1975.
- [36] R. A. Holley, S. Kusuoka, and D. W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.*, 83(2):333–347, 1989.
- [37] H. Huang and J. Qiu. On the mean-field limit for the consensus-based optimization. *Math. Methods Appl. Sci.*, 45(12):7814–7831, 2022.
- [38] H. Huang, J. Qiu, and K. Riedl. On the global convergence of particle swarm optimization methods. *Appl. Math. Optim.*, 88(2):Paper No. 30, 44, 2023.
- [39] P.-E. Jabin and Z. Wang. Mean field limit for stochastic particle systems. In *Active particles. Vol. 1. Advances in theory, models, and applications*, Model. Simul. Sci. Eng. Technol., pages 379–402. Birkhäuser/Springer, Cham, 2017.
- [40] D. Kalise, A. Sharma, and M. V. Tretyakov. Consensus-based optimization via jump-diffusion stochastic differential equations. *Math. Models Methods Appl. Sci.*, 33(2):289–339, 2023.
- [41] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, 2016.
- [42] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*, pages 1942–1948. IEEE, 1995.
- [43] J. Kim, M. Kang, D. Kim, S. Ha, and I. Yang. A stochastic consensus method for nonconvex optimization on the stiefel manifold. In *59th IEEE Conference on Decision and Control, CDC 2020, Jeju Island, South Korea, December 14-18, 2020*, pages 1050–1057. IEEE, 2020.
- [44] Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database, 2010.
- [45] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [46] J. Lu, E. Tadmor, and A. Zenginoglu. Swarm-based gradient descent method for non-convex optimization. *arXiv preprint arXiv:2211.17157*, 2022.
- [47] H. P. McKean, Jr. Propagation of chaos for a class of non-linear parabolic equations. In *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967)*, volume Session 7 of *Lecture Series in Differential Equations*, pages 41–57. Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Arlington, VA, 1967.

- [48] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018.
- [49] P. D. Miller. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.
- [50] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.*, 175(1-2, Ser. A):69–107, 2019.
- [51] B. Øksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, Berlin, sixth edition, 2003.
- [52] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.
- [53] E. Platen. An introduction to numerical methods for stochastic differential equations. In *Acta numerica, 1999*, volume 8 of *Acta Numer.*, pages 197–246. Cambridge Univ. Press, Cambridge, 1999.
- [54] L. A. Rastrigin. The convergence of the random search method in the extremal control of a many parameter system. *Automaton & Remote Control*, 24:1337–1342, 1963.
- [55] C. Reeves. Genetic algorithms. In *Handbook of metaheuristics*, volume 57 of *Internat. Ser. Oper. Res. Management Sci.*, pages 55–82. Kluwer Acad. Publ., Boston, MA, 2003.
- [56] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.
- [57] K. Riedl. Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law. *Eur. J. Appl. Math.*, page 32 pages, 2023.
- [58] K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier. Gradient is All You Need? How Consensus-Based Optimization can be Interpreted as a Stochastic Relaxation of Gradient Descent. *arXiv preprint arXiv:2306.09778*, 2023.
- [59] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
- [60] G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of artificial neural networks: an interacting particle system approach. *Comm. Pure Appl. Math.*, 75(9):1889–1935, 2022.
- [61] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: a law of large numbers. *SIAM J. Appl. Math.*, 80(2):725–752, 2020.
- [62] F. J. Solis and R. J.-B. Wets. Minimization by random search techniques. *Math. Oper. Res.*, 6(1):19–30, 1981.
- [63] A.-S. Sznitman. Topics in propagation of chaos. In *École d’Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Math.*, pages 165–251. Springer, Berlin, 1991.
- [64] C. Totzeck and M.-T. Wolfram. Consensus-based global optimization with personal best. *Math. Biosci. Eng.*, 17(5):6026–6044, 2020.

- [65] F. van den Bergh. *An analysis of particle swarm optimizers*. PhD thesis, University of Pretoria, 2007.
- [66] Y. Xu, Q. Lin, and T. Yang. Adaptive svrg methods under error bound conditions with unknown growth parameter. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Appendix: Extended proof details of Section 3.3

Extended proof of Lemma 15. By combining the ideas of [10, Lemma 3.4] with a Doob-like inequality, we derive a bound for $\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max \{ \|V_t^i\|_2^4, \|\bar{V}_t^i\|_2^4 \}$, which ensures that $\hat{\rho}_t^N, \bar{\rho}_t^N \in \mathcal{P}_4(\mathbb{R}^d)$ with high probability. Here, $\bar{\rho}^N$ denotes the empirical measure associated with the processes $(\bar{V}^i)_{i=1, \dots, N}$. For notational simplicity, but without loss of generality, we restrict ourselves to the case $H \equiv 1$ in what follows.

By employing the inequality $(x + y)^q \leq 2^{q-1}(x^q + y^q)$, $q \geq 1$ we note that

$$\begin{aligned} \|V_t^i\|_2^{2p} &\leq 2^{2p-1} \|V_0^i\|_2^{2p} + 2^{2(2p-1)} \lambda^{2p} \left\| \int_0^t (V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)) d\tau \right\|_2^{2p} \\ &\quad + 2^{2(2p-1)} \sigma^{2p} \left\| \int_0^t \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2 dB_\tau^i \right\|_2^{2p} \end{aligned}$$

for all $i = 1, \dots, N$. Taking first the supremum over $t \in [0, T]$ and consecutively the expectation on both sides of the former inequality yields

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \|V_t^i\|_2^{2p} &\leq 2^{2p-1} \mathbb{E} \|V_0^i\|_2^{2p} + 2^{2(2p-1)} \lambda^{2p} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t (V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)) d\tau \right\|_2^{2p} \\ &\quad + 2^{2(2p-1)} \sigma^{2p} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2 dB_\tau^i \right\|_2^{2p}. \end{aligned} \quad (68)$$

The second term on the right-hand side of (68) can be further bounded by

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t (V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)) d\tau \right\|_2^{2p} &\leq \max\{1, T^{2p-1}\} \mathbb{E} \sup_{t \in [0, T]} \int_0^t \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2^{2p} d\tau \\ &\leq \max\{1, T^{2p-1}\} \mathbb{E} \int_0^T \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2^{2p} d\tau \end{aligned} \quad (69)$$

as a consequence of Jensen's inequality. For the third term on the right-hand side of (68) we first note that the expression $\int_0^t \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2^2 dB_\tau^i$ is a martingale. This is due to [51, Corollary 3.2.6] since its expected quadratic variation is finite as required by [51, Definition 3.1.4]. The latter immediately follows from the regularity established in [10, Lemma 3.4]. Therefore we can apply the Burkholder-Davis-Gundy inequality [56, Chapter IV, Theorem 4.1], which gives for a generic constant C_{2p} depending only on the dimension d the bound

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \left\| \int_0^t \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2 dB_\tau^i \right\|_2^{2p} &\leq C_{2p} \sup_{t \in [0, T]} \mathbb{E} \left(\int_0^t \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2^2 d\tau \right)^p \\ &\leq C_{2p} \max\{1, T^{p-1}\} \mathbb{E} \int_0^T \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2^{2p} d\tau. \end{aligned} \quad (70)$$

Here, the latter step is again due to Jensen's inequality. The right-hand sides of (69) and (70) can be further bounded since

$$\begin{aligned} \mathbb{E} \int_0^T \|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2^{2p} d\tau &\leq 2^{2p-1} \mathbb{E} \int_0^T \left(\|V_\tau^i\|_2^{2p} + \|v_\alpha(\hat{\rho}_\tau^N)\|_2^{2p} \right) d\tau \\ &\leq 2^{2p-1} \mathbb{E} \int_0^T \left(\|V_\tau^i\|_2^{2p} + 2^{p-1} \left(b_1^p + b_2^p \int \|v\|_2^{2p} d\hat{\rho}_\tau^N(v) \right) \right) d\tau, \end{aligned} \quad (71)$$

where in the last step we made use of [10, Lemma 3.3], which shows that

$$\|v_\alpha(\hat{\rho}_\tau^N)\|_2^2 \leq \int \|v\|_2^2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\hat{\rho}_\tau^N)}} d\hat{\rho}_\tau^N(v) \leq b_1 + b_2 \int \|v\|_2^2 d\hat{\rho}_\tau^N(v),$$

with $b_1 = 0$ and $b_2 = e^{\alpha(\bar{\mathcal{E}} - \mathcal{E})}$ in the case that \mathcal{E} is bounded, and

$$b_1 = C_4^2 + b_2 \quad \text{and} \quad b_2 = 2 \frac{C_2}{C_3} \left(1 + \frac{1}{\alpha C_3} \frac{1}{C_4^2} \right) \quad (72)$$

in the case that \mathcal{E} satisfies the coercivity assumption (22). Inserting the upper bounds (69) and (70) together with the estimate (71) into (68) yields

$$\mathbb{E} \sup_{t \in [0, T]} \|V_t^i\|_2^{2p} \leq C \left(1 + \mathbb{E} \|V_0^i\|_2^{2p} + \mathbb{E} \int_0^T \|V_\tau^i\|_2^{2p} + \int \|v\|_2^{2p} d\widehat{\rho}_\tau^N(v) d\tau \right) \quad (73)$$

with a constant $C = C(p, \lambda, \sigma, d, T, b_1, b_2)$. Averaging (73) over i allows to bound

$$\begin{aligned} \mathbb{E} \sup_{t \in [0, T]} \int \|v\|_2^{2p} d\widehat{\rho}_t^N(v) &\leq C \left(1 + \mathbb{E} \int \|v\|_2^{2p} d\widehat{\rho}_0^N(v) + 2 \int_0^T \mathbb{E} \int \|v\|_2^{2p} d\widehat{\rho}_\tau^N(v) d\tau \right), \\ &\leq C \left(1 + \mathbb{E} \int \|v\|_2^{2p} d\widehat{\rho}_0^N(v) + 2 \int_0^T \mathbb{E} \sup_{\tau \in [0, \tau]} \int \|v\|_2^{2p} d\widehat{\rho}_\tau^N(v) d\tau \right), \end{aligned}$$

which ensures after an application of Grönwall's inequality, that $\mathbb{E} \sup_{t \in [0, T]} \int \|v\|_2^{2p} d\widehat{\rho}_t^N(v)$ is bounded independently of N provided $\rho_0 \in \mathcal{P}_{2p}(\mathbb{R}^d)$. Since this holds by the assumption $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ for $p = 2$, there exists a constant $K = K(\lambda, \sigma, d, T, b_1, b_2)$, in particular independently of N , such that $\mathbb{E} \sup_{t \in [0, T]} \int \|v\|_2^4 d\widehat{\rho}_t^N(v) \leq K$.

Following analogous arguments for \bar{V}_t^i allows to derive

$$\mathbb{E} \sup_{t \in [0, T]} \|\bar{V}_t^i\|_2^{2p} \leq C \left(1 + \mathbb{E} \|\bar{V}_0^i\|_2^{2p} + \mathbb{E} \int_0^T \|\bar{V}_\tau^i\|_2^{2p} + \int \|v\|_2^{2p} d\rho_\tau(v) d\tau \right) \quad (74)$$

in place of (73). Noticing that $\int \|v\|_2^{2p} d\rho_\tau(v) = \mathbb{E} \|\bar{V}_\tau^i\|_2^{2p}$ for all $\tau \in [0, T]$ and averaging the latter over i directly permits to prove $\mathbb{E} \sup_{t \in [0, T]} \int \|v\|_2^{2p} d\bar{\rho}_t(v) \leq K$ by applying Grönwall's inequality, again provided that $\rho_0 \in \mathcal{P}_{2p}(\mathbb{R}^d)$. With this being the case for $p = 2$ and by choosing K sufficiently large for either estimate, the statement follows from a union bound and Markov's inequality. More precisely,

$$\begin{aligned} \mathbb{P} \left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max \left\{ \|V_t^i\|_2^4, \|\bar{V}_t^i\|_2^4 \right\} > M \right) \\ \leq \mathbb{P} \left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|V_t^i\|_2^4 > M \right) + \mathbb{P} \left(\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|\bar{V}_t^i\|_2^4 > M \right) \\ \leq \frac{\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|V_t^i\|_2^4}{M} + \frac{\mathbb{E} \sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \|\bar{V}_t^i\|_2^4}{M} \leq 2 \frac{K}{M}. \end{aligned}$$

□

Extended proof of Proposition 16. By exploiting the boundedness of the dynamics established in Lemma 15 through a cutoff technique, we can follow the steps taken in [25, Theorem 3.1]. For notational simplicity, we restrict ourselves to the case $H \equiv 1$ in what follows. However, at the expense of minor technical modifications, the proof can be extended to the case of a Lipschitz-continuous active function H .

Let us define the cutoff function

$$I_M(t) = \begin{cases} 1, & \text{if } \frac{1}{N} \sum_{i=1}^N \max \left\{ \|V_\tau^i\|_2^4, \|\bar{V}_\tau^i\|_2^4 \right\} \leq M \text{ for all } \tau \in [0, t], \\ 0, & \text{else,} \end{cases} \quad (75)$$

which is adapted to the natural filtration and has the property $I_M(t) = I_M(t)I_M(\tau)$ for all $\tau \in [0, t]$. This allows to obtain for $\mathbb{E} \|V_t^i - \bar{V}_t^i\|_2^2 I_M(t)$ the inequality

$$\begin{aligned}
\mathbb{E} \|V_t^i - \bar{V}_t^i\|_2^2 I_M(t) &\leq 2\mathbb{E} \|V_0^i - \bar{V}_0^i\|_2^2 \\
&\quad + 4\lambda^2 \mathbb{E} \left\| \int_0^t ((V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)) - (\bar{V}_\tau^i - v_\alpha(\rho_\tau))) I_M(\tau) d\tau \right\|_2^2 \\
&\quad + 4\sigma^2 \mathbb{E} \left\| \int_0^t (\|V_\tau^i - v_\alpha(\hat{\rho}_\tau^N)\|_2 - \|\bar{V}_\tau^i - v_\alpha(\rho_\tau)\|_2) I_M(\tau) dB_\tau^i \right\|_2^2 \\
&\leq 2\mathbb{E} \|V_0^i - \bar{V}_0^i\|_2^2 \\
&\quad + 8\lambda^2 T \mathbb{E} \int_0^t (\|V_\tau^i - \bar{V}_\tau^i\|_2^2 + \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2) I_M(\tau) d\tau \\
&\quad + 8\sigma^2 d \mathbb{E} \int_0^t (\|V_\tau^i - \bar{V}_\tau^i\|_2^2 + \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2) I_M(\tau) d\tau,
\end{aligned}$$

where we used in the first step that the processes V_τ^i and \bar{V}_τ^i share the Brownian motion paths, and in the second step both Itô isometry and Jensen's inequality. Noting further that the processes also share the initial data, we are left with

$$\mathbb{E} \|V_t^i - \bar{V}_t^i\|_2^2 I_M(t) \leq 8(\lambda^2 T + \sigma^2 d) \int_0^t \mathbb{E} (\|V_\tau^i - \bar{V}_\tau^i\|_2^2 + \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2) I_M(\tau) d\tau, \quad (76)$$

where it remains to control $\mathbb{E} \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau)$. By means of Lemmas 25 and 26 below we have the bound

$$\begin{aligned}
\mathbb{E} \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau) &\leq 2\mathbb{E} \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\|_2^2 I_M(\tau) \\
&\quad + 2\mathbb{E} \|v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau) \\
&\leq C \left(\frac{1}{N} \sum_{i=1}^N \mathbb{E} \|V_\tau^i - \bar{V}_\tau^i\|_2^2 I_M(\tau) + N^{-1} \right) \\
&\leq C \left(\max_{i=1, \dots, N} \mathbb{E} \|V_\tau^i - \bar{V}_\tau^i\|_2^2 I_M(\tau) + N^{-1} \right)
\end{aligned} \quad (77)$$

for a constant $C = C(\alpha, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. After integrating the bound (77) into (76) and taking the maximum over i we are left with

$$\max_{i=1, \dots, N} \mathbb{E} \|V_t^i - \bar{V}_t^i\|_2^2 I_M(t) \leq C \int_0^t \max_{i=1, \dots, N} \mathbb{E} \|V_\tau^i - \bar{V}_\tau^i\|_2^2 I_M(\tau) d\tau + CTN^{-1}, \quad (78)$$

where C depends additionally on λ, σ, d and T , i.e., $C = C(\alpha, \lambda, \sigma, d, T, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. The second part of the statement now follows from an application of Grönwall's inequality and by noting that $\mathbb{1}_{\Omega_M} \leq I_M(t)$ pointwise and for all $t \in [0, T]$. \square

Lemma 25. *Let I_M be as defined in (75). Then, under the assumptions of Theorem 6, it holds*

$$\|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\|_2^2 I_M(\tau) \leq C \frac{1}{N} \sum_{i=1}^N \|V_\tau^i - \bar{V}_\tau^i\|_2^2 I_M(\tau)$$

for a constant $C = C(\alpha, C_1, C_2, M)$.

Proof. The proof follows the steps taken in [10, Lemmas 3.1 and 3.2].

Let us first note that by exploiting that the quantity $\frac{1}{N} \sum_{i=1}^N \|V_\tau^i\|_2^4$ is bounded uniformly by M due to the multiplication with $I_M(\tau)$, we obtain with Jensen's inequality that

$$\begin{aligned} \frac{e^{-\alpha \underline{\mathcal{E}}} I_M(\tau)}{\frac{1}{N} \sum_{i=1}^N \omega_\alpha(V_\tau^i)} &\leq \frac{I_M(\tau)}{\exp(-\alpha \frac{1}{N} \sum_{i=1}^N (\mathcal{E}(V_\tau^i) - \underline{\mathcal{E}}))} \leq \frac{I_M(\tau)}{\exp(-\alpha C_2 \frac{1}{N} \sum_{i=1}^N (1 + \|V_\tau^i\|_2^2))} \\ &\leq \exp(\alpha C_2 (1 + \sqrt{M})) =: c_M, \end{aligned} \quad (79)$$

where, in the second inequality, we used the assumption (21) on \mathcal{E} . An analogous statement can be obtained for the processes \bar{V}_τ^i .

For the norm of the difference between $v_\alpha(\hat{\rho}_\tau^N)$ and $v_\alpha(\bar{\rho}_\tau^N)$ we have the decomposition

$$\begin{aligned} \|v_\alpha(\hat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\|_2 I_M(\tau) &= \left\| \frac{\sum_{i=1}^N V_\tau^i \omega_\alpha(V_\tau^i)}{\sum_{j=1}^N \omega_\alpha(V_\tau^j)} - \frac{\sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} \right\|_2 I_M(\tau) \\ &\leq (\|T_1\|_2 + \|T_2\|_2 + \|T_3\|_2) I_M(\tau), \end{aligned} \quad (80)$$

where the terms T_1 , T_2 and T_3 are obtained by inserting mixed terms with respect to V_τ^i and \bar{V}_τ^i . They are defined implicitly below and their norm is bounded as follows. For the first term T_1 we have

$$\begin{aligned} \|T_1\|_2 I_M(\tau) &= \left\| \frac{1}{N} \sum_{i=1}^N (V_\tau^i - \bar{V}_\tau^i) \frac{\omega_\alpha(V_\tau^i)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right\|_2 I_M(\tau) \\ &\leq \frac{1}{N} \sum_{i=1}^N \|V_\tau^i - \bar{V}_\tau^i\|_2 \left| \frac{\omega_\alpha(V_\tau^i)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right| I_M(\tau) \\ &\leq \left| \frac{e^{-\alpha \underline{\mathcal{E}}} I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right| \frac{1}{N} \sum_{i=1}^N \|V_\tau^i - \bar{V}_\tau^i\|_2 I_M(\tau) \\ &\leq c_M \sqrt{\frac{1}{N} \sum_{i=1}^N \|V_\tau^i - \bar{V}_\tau^i\|_2^2} I_M(\tau), \end{aligned} \quad (81)$$

where we made use of (79) and Cauchy-Schwarz inequality in the last step. For the second term T_2 , by using the assumption (20) on \mathcal{E} in the third line and by following similar steps, we obtain

$$\begin{aligned} \|T_2\|_2 I_M(\tau) &= \left\| \frac{1}{N} \sum_{i=1}^N (\omega_\alpha(V_\tau^i) - \omega_\alpha(\bar{V}_\tau^i)) \frac{\bar{V}_\tau^i}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right\|_2 I_M(\tau) \\ &\leq \frac{1}{N} \sum_{i=1}^N |\omega_\alpha(V_\tau^i) - \omega_\alpha(\bar{V}_\tau^i)| \left\| \frac{\bar{V}_\tau^i}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right\|_2 I_M(\tau) \\ &\leq \alpha C_1 e^{-\alpha \underline{\mathcal{E}}} \frac{1}{N} \sum_{i=1}^N (\|V_\tau^i\|_2 + \|\bar{V}_\tau^i\|_2) \|V_\tau^i - \bar{V}_\tau^i\|_2 \frac{\|\bar{V}_\tau^i\|_2}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} I_M(\tau) \\ &\leq \frac{3}{2} \alpha C_1 \left| \frac{e^{-\alpha \underline{\mathcal{E}}} I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right| \sqrt{\frac{1}{N} \sum_{i=1}^N (\|V_\tau^i\|_2^4 + \|\bar{V}_\tau^i\|_2^4)} I_M(\tau) \\ &\quad \cdot \sqrt{\frac{1}{N} \sum_{i=1}^N \|V_\tau^i - \bar{V}_\tau^i\|_2^2} I_M(\tau) \\ &\leq 3\alpha C_1 c_M M^{\frac{1}{2}} \sqrt{\frac{1}{N} \sum_{i=1}^N \|V_\tau^i - \bar{V}_\tau^i\|_2^2} I_M(\tau). \end{aligned} \quad (82)$$

Analogously, for the third term T_3 , we get

$$\begin{aligned}
\|T_3\|_2 I_M(\tau) &= \left\| \frac{\sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} - \frac{\sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} \right\|_2 I_M(\tau) \\
&\leq \frac{1}{N} \sum_{j=1}^N |\omega_\alpha(\bar{V}_\tau^j) - \omega_\alpha(V_\tau^j)| \left\| \frac{\frac{1}{N} \sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i)}{\left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)\right) \left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)\right)} \right\|_2 I_M(\tau) \\
&\leq \alpha C_1 e^{-2\alpha\varepsilon} \frac{1}{N} \sum_{j=1}^N (\|V_\tau^j\|_2 + \|\bar{V}_\tau^j\|_2) \|V_\tau^j - \bar{V}_\tau^j\|_2 I_M(\tau) \\
&\quad \cdot \frac{\frac{1}{N} \sum_{i=1}^N \|\bar{V}_\tau^i\|_2 I_M(\tau)}{\left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha(V_\tau^j)\right) \left(\frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)\right)} \\
&\leq \sqrt{2} \alpha C_1 c_M^2 M^{\frac{1}{4}} \sqrt{\frac{1}{N} \sum_{j=1}^N (\|V_\tau^j\|_2^2 + \|\bar{V}_\tau^j\|_2^2)} I_M(\tau) \sqrt{\frac{1}{N} \sum_{j=1}^N \|V_\tau^j - \bar{V}_\tau^j\|_2^2} I_M(\tau) \\
&\leq 2\alpha C_1 c_M^2 M^{\frac{1}{2}} \sqrt{\frac{1}{N} \sum_{j=1}^N \|V_\tau^j - \bar{V}_\tau^j\|_2^2} I_M(\tau).
\end{aligned} \tag{83}$$

By inserting the three individual bounds (81), (82) and (83) into (80) and taking the squares of both sides, we obtain the upper bound from the statement. \square

Lemma 26. *Let $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and let I_M be as defined in (75). Then, under the assumptions of Theorem 6, it holds*

$$\sup_{\tau \in [0, T]} \mathbb{E} \|v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau) \leq CN^{-1} \tag{84}$$

for a constant $C = C(\alpha, C_2, M, \mathcal{M}_2, b_1, b_2)$, where \mathcal{M}_2 denotes the second-order moment bound of ρ and where b_1 and b_2 are the problem-dependent constants specified in (72).

Proof. The proof follows the steps taken in [25, Lemma 3.1].

By inserting a mixed term, we can bound the norm of the difference between $v_\alpha(\bar{\rho}_\tau^N)$ and $v_\alpha(\rho_\tau)$ by

$$\begin{aligned}
\|v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2 I_M(\tau) &= \left\| \sum_{i=1}^N \bar{V}_\tau^i \frac{\omega_\alpha(\bar{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} - \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\rho_\tau)}} d\rho_\tau(v) \right\|_2 I_M(\tau) \\
&\leq (\|T_1\|_2 + \|T_2\|_2) I_M(\tau),
\end{aligned} \tag{85}$$

where the terms T_1 and T_2 are defined implicitly and bounded in what follows. By utilizing the bound (79), for the first term T_1 , we get

$$\begin{aligned}
\|T_1\|_2 I_M(\tau) &= \left\| \sum_{i=1}^N \bar{V}_\tau^i \frac{\omega_\alpha(\bar{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} - \int v \frac{\omega_\alpha(v)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} d\rho_\tau(v) \right\|_2 I_M(\tau) \\
&= \left| \frac{I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} \right| \left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i) - \int v \omega_\alpha(v) d\rho_\tau(v) \right\|_2 \\
&\leq c_M e^{\alpha\varepsilon} \left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i) - \int v \omega_\alpha(v) d\rho_\tau(v) \right\|_2.
\end{aligned} \tag{86}$$

Similarly, for the second term we have

$$\begin{aligned}
\|T_2\|_2 I_M(\tau) &= \left\| \int v \frac{\omega_\alpha(v)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} d\rho_\tau(v) - \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\rho_\tau)}} d\rho_\tau(v) \right\|_2 I_M(\tau) \\
&= \left| \frac{I_M(\tau)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j)} \right| \|v_\alpha(\rho_\tau)\|_2 \left| \frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j) - \|\omega_\alpha\|_{L_1(\rho_\tau)} \right| \\
&\leq c_M e^{\alpha \varepsilon} \sqrt{b_1 + b_2 \mathcal{M}_2} \left| \frac{1}{N} \sum_{j=1}^N \omega_\alpha(\bar{V}_\tau^j) - \int \omega_\alpha(v) d\rho_\tau(v) \right|,
\end{aligned} \tag{87}$$

where the last step uses that by Jensen's inequality and [10, Lemma 3.3] it holds

$$\|v_\alpha(\rho_\tau)\|_2^2 \leq \int \|v\|_2^2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\rho_\tau)}} d\rho_\tau(v) \leq b_1 + b_2 \int \|v\|_2^2 d\rho_\tau(v) \leq b_1 + b_2 \mathcal{M}_2$$

with constants b_1 and b_2 as specified in (72) and \mathcal{M}_2 denoting a bound on the second-order moment of ρ , which exists according to the regularity of ρ established in Theorem 6 as a consequence of the initial regularity $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$. In order to further bound (86) and (87), respectively, let us introduce the random variables

$$Z_\tau^i = \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i) - \int v \omega_\alpha(v) d\rho_\tau(v) \quad \text{and} \quad z_\tau^i = \omega_\alpha(\bar{V}_\tau^i) - \int \omega_\alpha(v) d\rho_\tau(v),$$

which have zero expectation, i.e., $\mathbb{E} Z_\tau^i = 0$ and $\mathbb{E} z_\tau^i = 0$. Moreover, we observe that

$$\frac{1}{N} \sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i) - \int v \omega_\alpha(v) d\rho_\tau(v) = \frac{1}{N} \sum_{i=1}^N Z_\tau^i$$

and

$$\frac{1}{N} \sum_{i=1}^N \omega_\alpha(\bar{V}_\tau^i) - \int \omega_\alpha(v) d\rho_\tau(v) = \frac{1}{N} \sum_{i=1}^N z_\tau^i,$$

respectively. Moreover, due to the independence of the \bar{V}_τ^i 's the Z_τ^i 's are independent and thus satisfy $\mathbb{E} Z_\tau^i Z_\tau^j = 0$ for $i \neq j$. Using this we can rewrite

$$\begin{aligned}
\mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_\tau^i \omega_\alpha(\bar{V}_\tau^i) - \int v \omega_\alpha(v) d\rho_\tau(v) \right\|_2^2 &= \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N Z_\tau^i \right\|_2^2 = \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \sum_{j=1}^N \langle Z_\tau^i, Z_\tau^j \rangle \\
&= \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \|Z_\tau^i\|_2^2 = \frac{1}{N} \mathbb{E} \|Z_\tau^1\|_2^2 \leq 4e^{-\alpha \varepsilon} \mathcal{M}_2 \frac{1}{N},
\end{aligned} \tag{88}$$

where the inequality in the last step is due to the estimate

$$\begin{aligned}
\mathbb{E} \|Z_\tau^1\|_2^2 &\leq 2\mathbb{E} \|\bar{V}_\tau^1 \omega_\alpha(\bar{V}_\tau^1)\|_2^2 + 2 \left\| \int v \omega_\alpha(v) d\rho_\tau(v) \right\|_2^2 \\
&\leq 2e^{-\alpha \varepsilon} \left(\mathbb{E} \|\bar{V}_\tau^1\|_2^2 + \int \|v\|_2^2 d\rho_\tau(v) \right) \leq 4e^{-\alpha \varepsilon} \mathcal{M}_2.
\end{aligned}$$

Following analogous arguments and noting that

$$\mathbb{E} |z_\tau^1|^2 \leq 2\mathbb{E} |\omega_\alpha(\bar{V}_\tau^1)|^2 + 2 \left| \int \omega_\alpha(v) d\rho_\tau(v) \right|^2 \leq 4e^{-\alpha \varepsilon}$$

yields the inequality

$$\mathbb{E} \left| \frac{1}{N} \sum_{i=1}^N \omega_\alpha(\bar{V}_\tau^i) - \int \omega_\alpha(v) d\rho_\tau(v) \right|^2 = \frac{1}{N} \mathbb{E} |z_\tau^1|^2 \leq 4e^{-\alpha\varepsilon} \frac{1}{N}. \quad (89)$$

Taking the square and expectation on both sides of (86) and (87), inserting the two individual bounds (88) and (89), gives the statement after recalling (85). \square

License for [CBO-I].

The permission to reprint and include the material is printed on the next page(s).

Von: Kelly Thomas Thomas@siam.org
Betreff: RE: Request for Permission to use Material in my Dissertation (SIOPT M152780)
Datum: 14. März 2024 um 21:14
An: Konstantin Riedl konstantin.riedl@ma.tum.de

Dear Mr. Riedl:

SIAM is happy to give permission to reprint material from “Consensus-Based Optimization Methods Converge Globally“ (SIOPT M152780) and “Consensus-Based Optimization for Saddle Point Problems“ (SICON M154336) in your dissertation. Please acknowledge the original publications, using the complete bibliographic information if available.

Sincerely,

Kelly Thomas
Managing Editor
Society for Industrial and Applied Mathematics
3600 Market Street - 6th Floor
Philadelphia, PA 19104
thomas@siam.org / (267) 350-6387

Von: Konstantin Riedl konstantin.riedl@ma.tum.de
Betreff: Request for Permission to use Material in my Dissertation (SIOPT M152780)
Datum: 14. März 2024 um 18:02
An: Thomas@siam.org

Dear Mrs. Thomas,

As one of the authors of the article „Consensus-Based Optimization Methods Converge Globally“ (SIOPT M152780) which was submitted to the *SIAM Journal on Optimization*, I am reaching out to you as the Managing Editor of SIAM to request permission to include the paper in my dissertation (doctoral thesis).

Since I am pursuing a cumulative dissertation, it is necessary by the rules of my university, the Technical University of Munich, and the School of Computation, Information and Technology, that I provide and include in my dissertation a **written letter of approval from the publisher** for all my publications that are part of my dissertation.

I hereby kindly ask for such a confirmation from SIAM (by email or a weblink to your terms and conditions), which allows me to use the aforementioned article in my dissertation.

If you have any questions beforehand, do not hesitate to ask.

Best regards,
Konstantin

Konstantin Riedl, M.Sc.

Technical University of Munich
School of Computation, Information and Technology
Department of Mathematics
Chair for Applied Numerical Analysis

Munich Center for Machine Learning

Institute for Ethics in Artificial Intelligence

Email: konstantin.riedl@ma.tum.de

arXiv.org - Non-exclusive license to distribute

The URI <http://arxiv.org/licenses/nonexclusive-distrib/1.0/> is used to record the fact that the submitter granted the following license to arXiv.org on submission of an article:

- I grant arXiv.org a perpetual, non-exclusive license to distribute this article.
- I certify that I have the right to grant this license.
- I understand that submissions cannot be completely removed once accepted.
- I understand that arXiv.org reserves the right to reclassify or reject any submission.

Revision history

2004-01-16 - License above introduced as part of arXiv submission process

2007-06-21 - This HTML page created

[Contact](#)

Paper P2

Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law

M. Fornasier, T. Klock, and K. Riedl

*Applications of Evolutionary Computation - 25th European Conference,
EvoApplications 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22,
2022, Proceedings (2022)*

Paper Summary of [CBO-II]³⁵

In the paper “Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law,” presented at the *25th European Conference on the Applications of Evolutionary and Bio-Inspired Computation*, held as part of the *EvoStar Conference*, and published in the proceedings *Applications of Evolutionary Computation*, part of the *Lecture Notes in Computer Science* series, we prove global convergence in mean-field law for the continuous-time analog of the CBO method (2.2) in the setting of anisotropic noise.

CBO is a population-based metaheuristic derivative-free optimization method capable of globally minimizing nonconvex and nonsmooth functions, i.e., solving (2.1), in high dimensions, in particular, when using anisotropic noise as originally proposed and investigated in [Car+21]. It is based on stochastic swarm intelligence, and inspired by consensus dynamics and opinion formation. Compared to other metaheuristic algorithms like particle swarm optimization, CBO is of a simpler nature and, therefore, more amenable to a rigorous theoretical convergence analysis.

By adapting the analytical framework put forward in [CBO-I], we show in [CBO-II] that anisotropic CBO converges globally in mean-field law with a dimension-independent rate for a rich class of objectives under minimal assumptions on the initialization of the method [CBO-II, Theorem 2], see also Theorem 3.6. This is confirmed numerically in [CBO-II, Figure 1]. The relevance of such convergence result of the mean-field limit is demonstrated in this dissertation, where we combine it with a quantitative mean-field approximation result, see Proposition 3.16, and classical results of numerical approximation of SDEs in order to obtain probabilistic global convergence guarantees of the numerical CBO method, see Theorem 3.19. To motivate anisotropic CBO from a practical perspective and demonstrate empirically successful applications of the method in the high-dimensional setting already with limited computational capacities, we showcase numerical experiments on a benchmark problem from machine learning, which is well understood in the literature [CBO-II, Section 4]. More specifically, we train a shallow and a convolutional neural network classifier for the MNIST dataset of handwritten digits [LCB10]. For this, we efficiently implement the anisotropic CBO algorithm utilizing several tweaks in the implementation, such as random mini-batch ideas and a cooling strategy of the parameters as proposed in [Car+21; For+21].

KR’s Contributions. KR proposed to extend the analysis about the global convergence of CBO methods in mean-field law developed in [CBO-I] to CBO with anisotropic noise due to its superior performance for high-dimensional optimization problems. Together with TK, KR discussed the necessary modifications to the theory before working out the technical details, conducting the numerical experiments, in particular implementing CBO for training small neural networks using random mini-batch ideas, and writing a first draft of the paper, which was then discussed with, proofread by and refined together with TK and MF.

³⁵In this section, we follow [CBO-II, Abstract].

The following document is a reprint of

- [CBO-II] M. Fornasier, T. Klock, and K. Riedl. “Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law.” In: *Applications of Evolutionary Computation - 25th European Conference, EvoApplications 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings*. Ed. by J. L. J. Laredo, J. I. Hidalgo, and K. O. Babaagba. Vol. 13224. Lecture Notes in Computer Science. Springer, 2022, pp. 738–754.

The permission to reprint and include the material is provided after the reprint.

Note: The proof of [CBO-II, Theorem 2] in the print below contains a subtle technical mistake. A corrected version of it is available both on [arXiv](#) and at the end of [Section 3.1.1](#), where we prove [Theorem 3.6](#), which strengthens the original result [CBO-II, Theorem 2].



Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law

Massimo Fornasier^{1,2}, Timo Klock³, and Konstantin Riedl¹✉

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

`{massimo.fornasier,konstantin.riedl}@ma.tum.de`

² Munich Data Science Institute, Munich, Germany

³ Department of Numerical Analysis and Scientific Computing, Simula Research Laboratory, Oslo, Norway

`timo@simula.no`

Abstract. In this paper we study anisotropic consensus-based optimization (CBO), a population-based metaheuristic derivative-free optimization method capable of globally minimizing nonconvex and nonsmooth functions in high dimensions. CBO is based on stochastic swarm intelligence, and inspired by consensus dynamics and opinion formation. Compared to other metaheuristic algorithms like Particle Swarm Optimization, CBO is of a simpler nature and therefore more amenable to theoretical analysis. By adapting a recently established proof technique, we show that anisotropic CBO converges globally with a dimension-independent rate for a rich class of objective functions under minimal assumptions on the initialization of the method. Moreover, the proof technique reveals that CBO performs a convexification of the optimization problem as the number of particles goes to infinity, thus providing an insight into the internal CBO mechanisms responsible for the success of the method. To motivate anisotropic CBO from a practical perspective, we further test the method on a complicated high-dimensional benchmark problem, which is well understood in the machine learning literature.

Keywords: High-dimensional global optimization · Metaheuristics · Consensus-based optimization · Mean-field limit · Anisotropy

1 Introduction

Several problems arising throughout all quantitative disciplines are concerned with the global unconstrained optimization of a problem-dependent objective function $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ and the search for the associated minimizing argument

$$v^* = \arg \min_{v \in \mathbb{R}^d} \mathcal{E}(v),$$

which is assumed to exist and be unique in what follows. Because of nowadays data deluge such optimization problems are usually high-dimensional. In machine

learning, for instance, one is interested in finding the optimal parameters of a neural network (NN) to accomplish various tasks, such as clustering, classification, and regression. The availability of huge amounts of training data for various real-world applications allows practitioners to work with models involving a large number of trainable parameters aiming for a high expressivity and accuracy of the trained model. This makes the resulting optimization process a high-dimensional problem. Since typical model architectures consist of many layers with a large amount of neurons, and include nonlinear and potentially non-smooth activation functions, the training process is in general a high-dimensional nonconvex optimization problem and therefore a particularly hard task.

Metaheuristics have a long history as state-of-the-art methods when it comes to tackling hard optimization problems. Inspired by self-organization and collective behavior in nature or human society, such as the swarming of flocks of birds or schools of fish [3], or opinion formation [20], they orchestrate an interplay between locally confined procedures and global strategies, randomness and deterministic decisions, to ensure a robust search for the global minimizer. Some prominent examples are Random Search [19], Evolutionary Programming [7], Genetic Algorithms [11], Ant Colony Optimization [6], Particle Swarm Optimization [14] and Simulated Annealing [1].

CBO follows those guiding principles, but is of much simpler nature and more amenable to theoretical analysis. The method uses N particles V^1, \dots, V^N , which are initialized independently according to some law $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$, to explore the domain and to form a global consensus about the minimizer v^* as time passes. For parameters $\alpha, \lambda, \sigma > 0$ the dynamics of each particle is given by

$$dV_t^i = -\lambda (V_t^i - v_\alpha(\hat{\rho}_t^N)) dt + \sigma D(V_t^i - v_\alpha(\hat{\rho}_t^N)) dB_t^i, \tag{1}$$

where $\hat{\rho}_t^N$ denotes the empirical measure of the particles. The first term in (1) is a drift term dragging the respective particle towards the momentaneous consensus point, a weighted average of the particles' positions, computed as

$$v_\alpha(\hat{\rho}_t^N) := \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\hat{\rho}_t^N)}} d\hat{\rho}_t^N(v), \quad \text{with} \quad \omega_\alpha(v) := \exp(-\alpha \mathcal{E}(v))$$

and motivated by the fact that $v_\alpha(\hat{\rho}_t^N) \approx \arg \min_{i=1, \dots, N} \mathcal{E}(V_t^i)$ for $\alpha \gg 1$ if the arg min is unique. To feature the exploration of the energy landscape of \mathcal{E} , the second term in (1) is a diffusion injecting randomness into the dynamics through independent standard Brownian motions $((B_t^i)_{t \geq 0})_{i=1, \dots, N}$. The two commonly studied diffusion types are isotropic [2, 9, 18] and anisotropic [4] diffusion with

$$D(V_t^i - v_\alpha(\hat{\rho}_t^N)) = \begin{cases} \|V_t^i - v_\alpha(\hat{\rho}_t^N)\|_2 \text{Id}, & \text{for isotropic diffusion,} \\ \text{diag}(V_t^i - v_\alpha(\hat{\rho}_t^N)), & \text{for anisotropic diffusion,} \end{cases}$$

where $\text{Id} \in \mathbb{R}^{d \times d}$ is the identity matrix and $\text{diag} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ the operator mapping a vector onto a diagonal matrix with the vector as its diagonal. The term's scaling encourages in particular particles far from $v_\alpha(\hat{\rho}_t^N)$ to explore larger

regions. The coordinate-dependent scaling of anisotropic diffusion has proven to be particularly beneficial for high-dimensional optimization problems by yielding dimension-independent convergence rates (see Fig. 1) and therefore improving both computational complexity and success probability of the algorithm [4, 8].

A theoretical convergence analysis of the CBO dynamics is possible either on the microscopic level (1) or by analyzing the macroscopic behavior of the particle density through a mean-field limit. In the large particle limit a particle is not influenced by individual particles but only by the average behavior of all particles. As shown in [12], the empirical random particle measure $\hat{\rho}^N$ converges in law to the deterministic particle density $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, which weakly (see Definition 1) satisfies the non-linear Fokker-Planck equation

$$\partial_t \rho_t = \lambda \operatorname{div}((v - v_\alpha(\rho_t)) \rho_t) + \frac{\sigma^2}{2} \sum_{k=1}^d \partial_{kk} \left(D(v - v_\alpha(\rho_t))_{kk}^2 \rho_t \right). \quad (2)$$

A quantitative analysis of the convergence rate remains, on non-compact domains, an open problem, see, e.g., [9, Remark 2]. Analyzing a mean-field limit such as (2) allows for establishing strong qualitative theoretical guarantees about CBO methods, paving the way to understand the internal mechanisms at play.

Prior Arts. The original CBO work [18] proposes the dynamics (1) with isotropic diffusion, which is analyzed in the mean-field sense in [2]. Under a stringent well-preparedness condition about ρ_0 and \mathcal{C}^2 regularity of \mathcal{E} the authors show consensus formation of the particles at some \tilde{v} close to v^* by first establishing exponential decay of the variance $\operatorname{Var}(\rho_t)$ and consecutively showing $\tilde{v} \approx v^*$ as a consequence of the Laplace principle [17]. This analysis is extended to the anisotropic case in [4].

Motivated by the surprising phenomenon that, on average, individual particles of the CBO dynamics follow the gradient flow of $v \mapsto \|v - v^*\|_2^2$, see [9, Figure 1b], the authors of [9] develop a novel proof technique for showing global convergence of isotropic CBO in mean-field law to v^* under minimal assumptions. Following [9, Definition 1], we speak of convergence in mean-field law to v^* for the interacting particle dynamics (1), if the solution ρ_t to the associated mean-field limit dynamics (2) converges narrowly to the Dirac delta δ_{v^*} at v^* for $t \rightarrow \infty$. The proof is based on showing an exponential decay of the energy functional $\mathcal{V} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}_{\geq 0}$, given by

$$\mathcal{V}(\rho_t) := \frac{1}{2} \int \|v - v^*\|_2^2 d\rho_t(v). \quad (3)$$

This simultaneously ensures consensus formation and convergence of ρ_t to δ_{v^*} .

Contribution. In view of the effectiveness and efficiency of CBO methods with anisotropic diffusion for high-dimensional optimization problems, a thorough understanding is of considerable interest. As we illustrate in Fig. 1, anisotropic

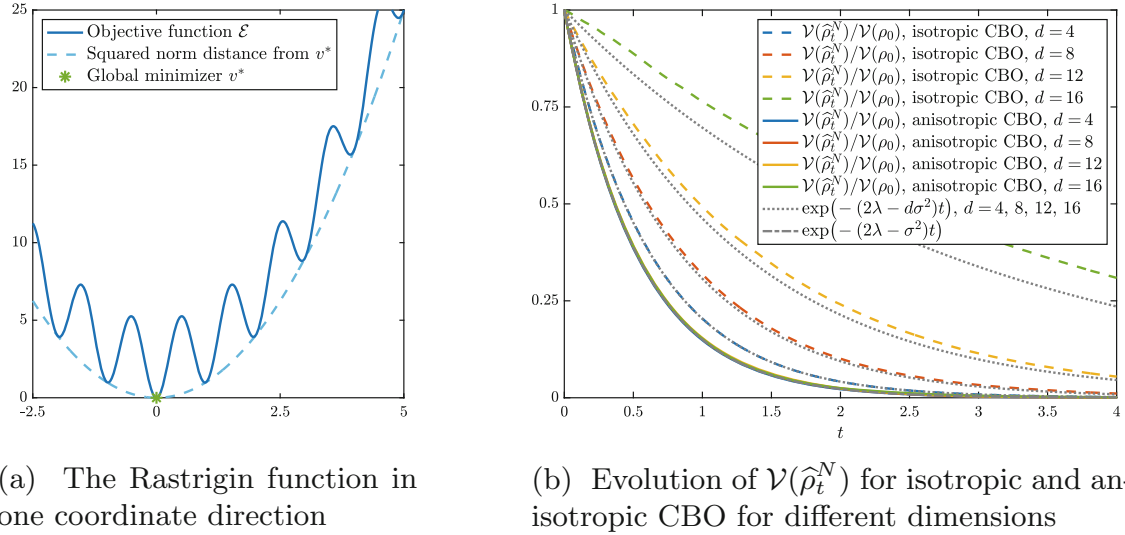


Fig. 1. A demonstration of the benefit of using anisotropic diffusion in CBO. For the Rastrigin function $\mathcal{E}(v) = \sum_{k=1}^d v_k^2 + \frac{5}{2}(1 - \cos(2\pi v_k))$ with $v^* = 0$ and spurious local minima (see (a)), we evolve the discretized system of isotropic and anisotropic CBO using $N = 320000$ particles, discrete time step size $\Delta t = 0.01$ and $\alpha = 10^{15}$, $\lambda = 1$, and $\sigma = 0.32$ for different dimensions $d \in \{4, 8, 12, 16\}$. We observe in (b) that the convergence rate of the energy functional $\mathcal{V}(\hat{\rho}_t^N)$ for isotropic CBO (dashed lines) is affected by the ambient dimension d , whereas anisotropic CBO (solid lines) converges independently from d with rate $(2\lambda - \sigma^2)$.

CBO [4] converges with a dimension-independent rate as opposed to isotropic CBO [2, 9, 18], making it a particularly interesting choice for problems in high-dimensional spaces, e.g., from signal processing and machine learning applications. In this work we extend the analysis of [9] from isotropic CBO to CBO with anisotropic diffusion. More precisely, we show global convergence of the anisotropic CBO dynamics in mean-field law to the global minimizer v^* under minimal assumptions about the initial measure ρ_0 and for a rich class of objectives \mathcal{E} . Furthermore, utilizing some tweaks in the implementation of anisotropic CBO, such as a random mini-batch idea and a cooling strategy of the parameters as proposed in [4, 10], we show that CBO performs well, in fact, almost on par with state-of-the-art gradient-based methods, on a long-studied machine learning benchmark in 2000 dimensions, despite using just 100 particles and no gradient information. This encourages the use and further investigation of CBO as a training algorithm for challenging machine learning tasks.

Organization. In Sect. 2 we first recall details about the well-posedness of the mean-field dynamics (2) in the case of anisotropic diffusion before we present the main theoretical result about the convergence of anisotropic CBO in mean-field law. The proof follows in Sect. 3. Section 4 illustrates the practicability of the method on a benchmark problem and Sect. 5 concludes the paper.

For the sake of reproducible research, in the GitHub repository <https://github.com/KonstantinRiedl/CBOGlobalConvergenceAnalysis> we provide the Matlab code implementing the CBO algorithm used in this work.

Notation. $B_r^\infty(u)$ is a closed ℓ^∞ ball in \mathbb{R}^d with center u and radius r . For the space of continuous functions $f : X \rightarrow Y$ we write $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$, $n \in \mathbb{N}$, and a suitable topological space Y . For $X \subset \mathbb{R}^n$ open and for $Y = \mathbb{R}^m$, $m \in \mathbb{N}$, the function space $\mathcal{C}_c^k(X, Y)$ contains functions $f \in \mathcal{C}(X, Y)$ that are k -times continuously differentiable and compactly supported. Y is omitted if $Y = \mathbb{R}$.

The objects of study are laws of stochastic processes, $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, where $\mathcal{P}(\mathbb{R}^d)$ contains all Borel probability measures over \mathbb{R}^d . $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ is a snapshot of such law at time t and ϱ some fixed distribution. Measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite p -th moment are collected in $\mathcal{P}_p(\mathbb{R}^d)$. For any $1 \leq p < \infty$, W_p denotes the Wasserstein- p distance. $\mathbb{E}(\varrho)$ is the expectation of a probability measure ϱ .

2 Global Convergence in Mean-Field Law

In this section we first recite a well-posedness result about the Fokker-Planck Eq. (2) and then present the main result about global convergence.

2.1 Definition of Weak Solutions and Well-Posedness

We begin by defining weak solutions of the Fokker-Planck Eq. (2).

Definition 1. Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$, $T > 0$. We say $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ satisfies the Fokker-Planck Eq. (2) with initial condition ρ_0 in the weak sense in the time interval $[0, T]$, if we have for all $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and all $t \in (0, T)$

$$\begin{aligned} \frac{d}{dt} \int \phi(v) d\rho_t(v) &= -\lambda \int \sum_{k=1}^d (v - v_\alpha(\rho_t))_k \partial_k \phi(v) d\rho_t(v) \\ &+ \frac{\sigma^2}{2} \int \sum_{k=1}^d D(V_t^i - v_\alpha(\widehat{\rho}_t^N))_{kk}^2 \partial_{kk}^2 \phi(v) d\rho_t(v) \end{aligned} \tag{4}$$

and $\lim_{t \rightarrow 0} \rho_t = \rho_0$ pointwise.

In what follows the case of CBO with anisotropic diffusion is considered, i.e., $D(V_t^i - v_\alpha(\widehat{\rho}_t^N)) = \text{diag}(V_t^i - v_\alpha(\widehat{\rho}_t^N))$ in Eqs. (1), (2) and (4).

Analogously to the well-posedness results [2, Theorems 3.1, 3.2] for CBO with isotropic diffusion, we can obtain well-posedness of (2) for anisotropic CBO.

Theorem 1. Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and consider $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ with $\underline{\mathcal{E}} := \mathcal{E}(v^*) > -\infty$, which, for some constants $C_1, C_2 > 0$, satisfies

$$\begin{aligned} |\mathcal{E}(v) - \mathcal{E}(w)| &\leq C_1(\|v\|_2 + \|w\|_2) \|v - w\|_2, \quad \text{for all } v, w \in \mathbb{R}^d, \\ \mathcal{E}(v) - \underline{\mathcal{E}} &\leq C_2(1 + \|v\|_2^2), \quad \text{for all } v \in \mathbb{R}^d. \end{aligned}$$

If in addition, either $\sup_{v \in \mathbb{R}^d} \mathcal{E}(v) < \infty$, or, for some $C_3, C_4 > 0$, \mathcal{E} satisfies

$$\mathcal{E}(v) - \underline{\mathcal{E}} \geq C_3 \|v\|_2^2, \quad \text{for all } \|v\|_2 \geq C_4,$$

then there exists a law $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ weakly satisfying Eq. (2).

Proof. The proof is based on the Leray-Schauder fixed point theorem and uses the same arguments as the ones provided for [2, Theorems 3.1, 3.2].

Remark 1. As discussed in [9, Remark 7], the proof of Theorem 1 justifies an extension of the test function space $\mathcal{C}_c^\infty(\mathbb{R}^d)$ in Definition 1 to

$$\begin{aligned} \mathcal{C}_*^2(\mathbb{R}^d) := \{ \phi \in \mathcal{C}^2(\mathbb{R}^d) : |\partial_k \phi(v)| \leq c(1 + |v_k|) \text{ and } \|\partial_{kk}^2 \phi\|_\infty < \infty \\ \text{for all } k \in \{1, \dots, d\} \text{ and some constant } c > 0 \}. \end{aligned}$$

2.2 Main Results

We now present the main result about global convergence in mean-field law for objective functions that satisfy the following conditions.

Definition 2 (Assumptions). We consider functions $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$, for which

A1 there exists $v^* \in \mathbb{R}^d$ such that $\mathcal{E}(v^*) = \inf_{v \in \mathbb{R}^d} \mathcal{E}(v) =: \underline{\mathcal{E}}$, and

A2 there exist $\mathcal{E}_\infty, R_0, \eta > 0$, and $\nu \in (0, \infty)$ such that

$$\|v - v^*\|_\infty \leq \frac{1}{\eta} (\mathcal{E}(v) - \underline{\mathcal{E}})^\nu \quad \text{for all } v \in B_{R_0}^\infty(v^*), \tag{5}$$

$$\mathcal{E}_\infty < \mathcal{E}(v) - \underline{\mathcal{E}} \quad \text{for all } v \in (B_{R_0}^\infty(v^*))^c. \tag{6}$$

Assumption A2 can be regarded as a tractability condition of the energy landscape around the minimizer and in the farfield. Equation (5) requires the local coercivity of \mathcal{E} , whereas (6) prevents that $\mathcal{E}(v) \approx \underline{\mathcal{E}}$ far away from v^* .

Definition 2 covers a wide range of function classes, including for instance the Rastrigin function, see Fig. 1a, and objectives related to various machine learning tasks, see, e.g., [10].

Theorem 2. Let \mathcal{E} be as in Definition 2. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ be such that

$$\rho_0(B_r^\infty(v^*)) > 0 \quad \text{for all } r > 0. \tag{7}$$

Define $\mathcal{V}(\rho_t) := 1/2 \int \|v - v^*\|_2^2 d\rho_t(v)$. Fix any $\varepsilon \in (0, \mathcal{V}(\rho_0))$ and $\tau \in (0, 1)$, parameters $\lambda, \sigma > 0$ with $2\lambda > \sigma^2$, and the time horizon

$$T^* := \frac{1}{(1 - \tau)(2\lambda - \sigma^2)} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right). \tag{8}$$

Then there exists $\alpha_0 > 0$, which depends (among problem dependent quantities) on ε and τ , such that for all $\alpha > \alpha_0$, if $\rho \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d))$ is a weak solution

to the Fokker-Planck Eq. (2) on the time interval $[0, T^*]$ with initial condition ρ_0 , we have $\min_{t \in [0, T^*]} \mathcal{V}(t) \leq \varepsilon$. Furthermore, until $\mathcal{V}(\rho_t)$ reaches the prescribed accuracy ε , we have the exponential decay

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0) \exp(-(1 - \tau)(2\lambda - \sigma^2)t)$$

and, up to a constant, the same behavior for $W_2^2(\rho_t, \delta_{v^*})$.

The rate of convergence $(2\lambda - \sigma^2)$ obtained in Theorem 2 is confirmed numerically by the experiments depicted in Fig. 1. We emphasize the dimension-independent convergence of CBO with anisotropic diffusion, contrasting the dimension-dependent rate $(2\lambda - d\sigma^2)$ of isotropic CBO, cf. [9, Theorem 12].

3 Proof of Theorem 2

This section provides the proof details for Theorem 2, starting with a sketch in Sect. 3.1. Sections 3.2–3.4 present statements, which are needed in the proof and may be of independent interest. Section 3.5 completes the proof.

Remark 2. Without loss of generality we assume $\underline{\mathcal{E}} = 0$ throughout the proof.

3.1 Proof Sketch

The main idea is to show that $\mathcal{V}(\rho_t)$ satisfies the differential inequality

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -(1 - \tau)(2\lambda - \sigma^2) \mathcal{V}(\rho_t) \quad (9)$$

until $\mathcal{V}(\rho_T) \leq \varepsilon$. The first step towards (9) is to derive a differential inequality for $\mathcal{V}(\rho_t)$ using the dynamics of ρ , which is done in Lemma 1. In order to control the appearing quantity $\|v_\alpha(\rho_t) - v^*\|_2$, we establish a quantitative Laplace principle. Namely, under the inverse continuity property A2, Proposition 1 shows

$$\|v_\alpha(\rho_t) - v^*\|_2 \lesssim \ell(r) + \frac{\sqrt{d} \exp(-\alpha r)}{\rho_t(B_r^\infty(v^*))}, \quad \text{for sufficiently small } r > 0,$$

where ℓ is decreasing with $\ell(r) \rightarrow 0^+$ as $r \rightarrow 0$. Thus, $\|v_\alpha(\rho_t) - v^*\|_2$ can be made arbitrarily small by suitable choices of $r \ll 1$ and $\alpha \gg 1$, as long as we can guarantee $\rho_t(B_r^\infty(v^*)) > 0$ for all $r > 0$ and at all times $t \in [0, T]$. The latter requires non-zero initial mass $\rho_0(B_r^\infty(v^*))$ as well as an active Brownian motion, as made rigorous in Proposition 2.

3.2 Evolution of the Mean-Field Limit

We now derive the evolution inequality of the energy functional $\mathcal{V}(\rho_t)$.

Lemma 1. *Let $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, and fix $\alpha, \lambda, \sigma > 0$. Moreover, let $T > 0$ and let $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d))$ be a weak solution to Eq. (2). Then $\mathcal{V}(\rho_t)$ satisfies*

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) \leq & - (2\lambda - \sigma^2) \mathcal{V}(\rho_t) + \sqrt{2} (\lambda + \sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2 \\ & + \frac{\sigma^2}{2} \|v_\alpha(\rho_t) - v^*\|_2^2. \end{aligned}$$

Proof. Noting that $\phi(v) = 1/2 \|v - v^*\|_2^2$ is in $\mathcal{C}_*^2(\mathbb{R}^d)$ and recalling that ρ satisfies the identity (4) for all test functions in $\mathcal{C}_*^2(\mathbb{R}^d)$, see Remark 1, we obtain

$$\frac{d}{dt} \mathcal{V}(\rho_t) = -\lambda \int \langle v - v^*, v - v_\alpha(\rho_t) \rangle d\rho_t(v) + \frac{\sigma^2}{2} \int \|v - v_\alpha(\rho_t)\|_2^2 d\rho_t(v),$$

where we used $\partial_k \phi(v) = (v - v^*)_k$ and $\partial_{kk}^2 \phi(v) = 1$ for all $k \in \{1, \dots, d\}$. Following the steps taken in [9, Lemma 14] yields the statement. \square

3.3 Quantitative Laplace Principle

The Laplace principle asserts that $-\log(\|\omega_\alpha\|_{L_1(\varrho)})/\alpha \rightarrow \underline{\mathcal{E}}$ as $\alpha \rightarrow \infty$ as long as the global minimizer v^* is in the support of ϱ . Under the assumption of the inverse continuity property this can be used to qualitatively characterize the proximity of $v_\alpha(\varrho)$ to the global minimizer v^* . However, as it neither allows to quantify this proximity nor gives a suggestion on how to choose α to reach a certain approximation quality, we introduced a quantitative version in [9, Proposition 17], which we now adapt suitably to satisfy the anisotropic setting.

Proposition 1. *Let $\underline{\mathcal{E}} = 0$, $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$ we define $\mathcal{E}_r := \sup_{v \in B_r^\infty(v^*)} \mathcal{E}(v)$. Then, under the inverse continuity property A2, for any $r \in (0, R_0]$ and $q > 0$ such that $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, we have*

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \frac{\sqrt{d}(q + \mathcal{E}_r)^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q)}{\varrho(B_r^\infty(v^*))} \int \|v - v^*\|_2 d\varrho(v).$$

Proof. Following the lines of the proof of [9, Proposition 17] but replacing all ℓ^2 balls and norms by ℓ^∞ balls and norms, respectively, we obtain

$$\|v_\alpha(\varrho) - v^*\|_\infty \leq \frac{(q + \mathcal{E}_r)^\nu}{\eta} + \frac{\exp(-\alpha q)}{\varrho(B_r^\infty(v^*))} \int \|v - v^*\|_\infty d\varrho(v).$$

The statement now follows noting that $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \sqrt{d} \|\cdot\|_\infty$. \square

3.4 A Lower Bound for the Probability Mass Around v^*

In this section we provide a lower bound for the probability mass of $\rho_t(B_r^\infty(v^*))$, where $r > 0$ is a small radius. This is achieved by defining a mollifier ϕ_r so that $\rho_t(B_r^\infty(v^*)) \geq \int \phi_r(v) d\rho_t(v)$ and studying the evolution of the right-hand side.

Lemma 2. For $r > 0$ we define the mollifier $\phi_r : \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$\phi_r(v) := \begin{cases} \prod_{k=1}^d \exp\left(1 - \frac{r^2}{r^2 - (v - v^*)_k^2}\right), & \text{if } \|v - v^*\|_\infty < r, \\ 0, & \text{else.} \end{cases} \tag{10}$$

We have $\text{Im}(\phi_r) = [0, 1]$, $\text{supp}(\phi_r) = B_r^\infty(v^*)$, $\phi_r \in C_c^\infty(\mathbb{R}^d)$ and

$$\begin{aligned} \partial_k \phi_r(v) &= -2r^2 \frac{(v - v^*)_k}{\left(r^2 - (v - v^*)_k^2\right)^2} \phi_r(v), \\ \partial_{kk}^2 \phi_r(v) &= 2r^2 \left(\frac{2 \left(2(v - v^*)_k^2 - r^2\right) (v - v^*)_k^2 - \left(r^2 - (v - v^*)_k^2\right)^2}{\left(r^2 - (v - v^*)_k^2\right)^4} \right) \phi_r(v). \end{aligned}$$

Proof. ϕ_r is a tensor product of classical well-studied mollifiers. □

Proposition 2. Let $T > 0$, $r > 0$, and fix parameters $\alpha, \lambda, \sigma > 0$. Assume $\rho \in C([0, T], \mathcal{P}(\mathbb{R}^d))$ weakly solves the Fokker-Planck Eq. (2) in the sense of Definition 1 with initial condition $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ and for $t \in [0, T]$. Furthermore, denote $B := \sup_{t \in [0, T]} \|v_\alpha(\rho_t) - v^*\|_\infty$. Then, for all $t \in [0, T]$ we have

$$\begin{aligned} \rho_t(B_r^\infty(v^*)) &\geq \left(\int \phi_r(v) d\rho_0(v) \right) \exp(-qt), \\ \text{for } q &:= 2d \max \left\{ \frac{\lambda(cr + B\sqrt{c})}{(1-c)^2 r} + \frac{\sigma^2(cr^2 + B^2)(2c+1)}{(1-c)^4 r^2}, \frac{2\lambda^2}{(2c-1)\sigma^2} \right\}, \end{aligned} \tag{11}$$

where $c \in (1/2, 1)$ can be any constant that satisfies $(1 - c)^2 \leq (2c - 1)c$.

Remark 3. In order to ensure a finite decay rate $q < \infty$ in Proposition 2 it is crucial to have a non-vanishing diffusion $\sigma > 0$.

Proof (Proposition 2). By the properties of the mollifier in Lemma 2 we have

$$\rho_t(B_r^\infty(v^*)) \geq \int \phi_r(v) d\rho_t(v).$$

Our strategy is to derive a lower bound for the right-hand side of this inequality. Using the weak solution property of ρ and the fact that $\phi_r \in C_c^\infty(\mathbb{R}^d)$, we obtain

$$\begin{aligned} \frac{d}{dt} \int \phi_r(v) d\rho_t(v) &= \sum_{k=1}^d \int (T_{1k}(v) + T_{2k}(v)) d\rho_t(v) \tag{12} \\ \text{with } T_{1k}(v) &:= -\lambda (v - v_\alpha(\rho_t))_k \partial_k \phi_r(v) \\ \text{and } T_{2k}(v) &:= \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 \partial_{kk}^2 \phi_r(v) \end{aligned}$$

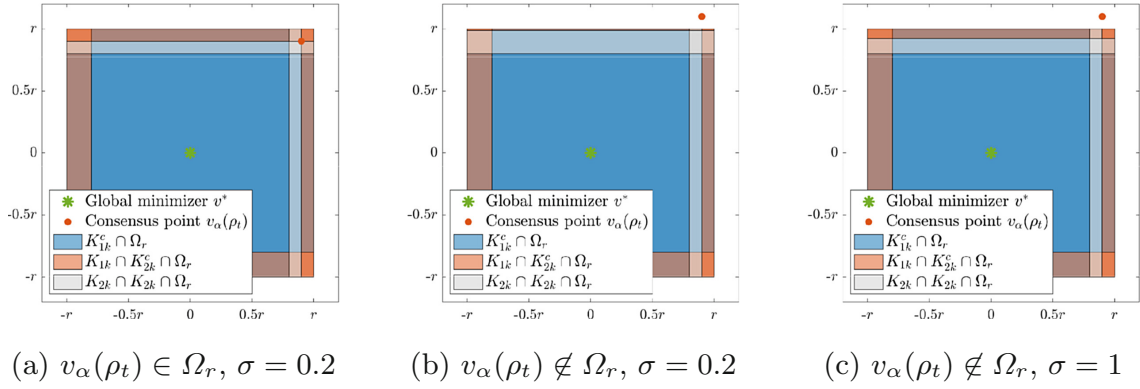


Fig. 2. Visualization of the decomposition of Ω_r as in (15) for different positions of $v_\alpha(\rho_t)$ and values of σ .

for $k \in \{1, \dots, d\}$. We now aim for showing $T_{1k}(v) + T_{2k}(v) \geq -q\phi_r(v)$ uniformly on \mathbb{R}^d individually for each k and for $q > 0$ as in the statement. Since the mollifier ϕ_r and its derivatives vanish outside of $\Omega_r := \{v \in \mathbb{R}^d : \|v - v^*\|_\infty < r\}$ we restrict our attention to the open ℓ^∞ -ball Ω_r . To achieve the lower bound over Ω_r , we introduce for each $k \in \{1, \dots, d\}$ the subsets

$$K_{1k} := \{v \in \mathbb{R}^d : |(v - v^*)_k| > \sqrt{cr}\} \tag{13}$$

and

$$K_{2k} := \left\{ v \in \mathbb{R}^d : -\lambda (v - v_\alpha(\rho_t))_k (v - v^*)_k \left(r^2 - (v - v^*)_k^2 \right)^2 > \tilde{c}r^2 \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 (v - v^*)_k^2 \right\}, \tag{14}$$

where $\tilde{c} := 2c - 1 \in (0, 1)$. For fixed k we now decompose Ω_r according to

$$\Omega_r = (K_{1k}^c \cap \Omega_r) \cup (K_{1k} \cap K_{2k}^c \cap \Omega_r) \cup (K_{1k} \cap K_{2k} \cap \Omega_r), \tag{15}$$

which is illustrated in Fig. 2 for different positions of $v_\alpha(\rho_t)$ and values of σ . In the following we treat each of these three subsets separately.

Subset $K_{1k}^c \cap \Omega_r$: We have $|(v - v^*)_k| \leq \sqrt{cr}$ for each $v \in K_{1k}^c$, which can be used to independently derive lower bounds for both terms T_{1k} and T_{2k} . For T_{1k} , we insert the expression for $\partial_k \phi_r$ from Lemma 2 to get

$$\begin{aligned} T_{1k}(v) &= 2r^2 \lambda (v - v_\alpha(\rho_t))_k \frac{(v - v^*)_k}{\left(r^2 - (v - v^*)_k^2 \right)^2} \phi_r(v) \\ &\geq -2r^2 \lambda \frac{|(v - v_\alpha(\rho_t))_k| |(v - v^*)_k|}{\left(r^2 - (v - v^*)_k^2 \right)^2} \phi_r(v) \geq -\frac{2\lambda(\sqrt{cr} + B)\sqrt{c}}{(1 - c)^2 r} \phi_r(v) \\ &=: -q_1 \phi_r(v), \end{aligned}$$

where $|(v - v_\alpha(\rho_t))_k| \leq |(v - v^*)_k| + |(v^* - v_\alpha(\rho_t))_k| \leq \sqrt{cr} + B$ is used in the last inequality. For T_2 we insert the expression for $\partial_{kk}^2 \phi_r$ from Lemma 2 to obtain

$$T_{2k}(v) = \sigma^2 r^2 (v - v_\alpha(\rho_t))_k^2 \frac{2 \left(2(v - v^*)_k^2 - r^2 \right) (v - v^*)_k^2 - \left(r^2 - (v - v^*)_k^2 \right)^2}{\left(r^2 - (v - v^*)_k^2 \right)^4} \phi_r(v) \\ \geq -\frac{2\sigma^2(cr^2 + B^2)(2c + 1)}{(1 - c)^4 r^2} \phi_r(v) =: -q_2 \phi_r(v),$$

where the last inequality uses $(v - v_\alpha(\rho_t))_k^2 \leq (\sqrt{cr} + B)^2 \leq 2(cr^2 + B^2)$.

Subset $K_{1k} \cap K_{2k}^c \cap \Omega_r$: As $v \in K_{1k}$ we have $|(v - v^*)_k|_2 > \sqrt{cr}$. We observe that $T_{1k}(v) + T_{2k}(v) \geq 0$ for all v in this subset whenever

$$\left(-\lambda (v - v_\alpha(\rho_t))_k (v - v^*)_k + \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 \right) \left(r^2 - (v - v^*)_k^2 \right)^2 \\ \leq \sigma^2 (v - v_\alpha(\rho_t))_k^2 \left(2(v - v^*)_k^2 - r^2 \right) (v - v^*)_k^2. \tag{16}$$

The first term on the left-hand side in (16) can be bounded from above exploiting that $v \in K_{2k}^c$ and by using the relation $\tilde{c} = 2c - 1$. More precisely, we have

$$-\lambda (v - v_\alpha(\rho_t))_k (v - v^*)_k \left(r^2 - (v - v^*)_k^2 \right)^2 \leq \tilde{c} r^2 \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 (v - v^*)_k^2 \\ = (2c - 1) r^2 \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 (v - v^*)_k^2 \leq \left(2(v - v^*)_k^2 - r^2 \right) \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 (v - v^*)_k^2,$$

where the last inequality follows since $v \in K_{1k}$. For the second term on the left-hand side in (16) we can use $(1 - c)^2 \leq (2c - 1)c$ as per assumption, to get

$$\frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 \left(r^2 - (v - v^*)_k^2 \right)^2 \leq \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 (1 - c)^2 r^4 \\ \leq \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 (2c - 1) r^2 c r^2 \leq \frac{\sigma^2}{2} (v - v_\alpha(\rho_t))_k^2 \left(2(v - v^*)_k^2 - r^2 \right) (v - v^*)_k^2.$$

Hence, (16) holds and we have $T_{1k}(v) + T_{2k}(v) \geq 0$ uniformly on this subset.

Subset $K_{1k} \cap K_{2k} \cap \Omega_r$: As $v \in K_{1k}$ we have $|(v - v^*)_k|_2 > \sqrt{cr}$. We first note that $T_{1k}(v) = 0$ whenever $\sigma^2 (v - v_\alpha(\rho_t))_k^2 = 0$, provided $\sigma > 0$, so nothing needs to be done if $v_k = (v_\alpha(\rho_t))_k$. Otherwise, if $\sigma^2 (v - v_\alpha(\rho_t))_k^2 > 0$, we exploit $v \in K_{2k}$ to get

$$\frac{(v - v_\alpha(\rho_t))_k (v - v^*)_k}{\left(r^2 - (v - v^*)_k^2 \right)^2} \geq \frac{-|(v - v_\alpha(\rho_t))_k| |(v - v^*)_k|}{\left(r^2 - (v - v^*)_k^2 \right)^2} \\ > \frac{2\lambda (v - v_\alpha(\rho_t))_k (v - v^*)_k}{\tilde{c} r^2 \sigma^2 |(v - v_\alpha(\rho_t))_k| |(v - v^*)_k|} \geq -\frac{2\lambda}{\tilde{c} r^2 \sigma^2}.$$

Using this, T_{1k} can be bounded from below by

$$T_{1k}(v) = 2r^2\lambda(v - v_\alpha(\rho_t))_k \frac{(v - v^*)_k}{\left(r^2 - (v - v^*)_k\right)^2} \phi_r(v) \geq -\frac{4\lambda^2}{\tilde{c}\sigma^2} \phi_r(v) =: -q_3\phi_r(v).$$

For T_{2k} , the nonnegativity of $\sigma^2(v - v_\alpha(\rho_t))_k^2$ implies $T_{2k}(v) \geq 0$, whenever

$$2\left(2(v - v^*)_k^2 - r^2\right)(v - v^*)_k \geq \left(r^2 - (v - v^*)_k\right)^2.$$

This holds for $v \in K_{1k}$, if $2(2c - 1)c \geq (1 - c)^2$ as implied by the assumption.

Concluding the Proof: Using the evolution of ϕ_r as in (12) and the individual decompositions of Ω_r for the terms $T_{1k} + T_{2k}$, we now get

$$\begin{aligned} \frac{d}{dt} \int \phi_r(v) d\rho_t(v) &= \sum_{k=1}^d \left(\int_{K_{1k} \cap K_{2k}^c \cap \Omega_r} \underbrace{(T_{1k}(v) + T_{2k}(v))}_{\geq 0} d\rho_t(v) \right. \\ &\quad \left. + \int_{K_{1k} \cap K_{2k} \cap \Omega_r} \underbrace{(T_{1k}(v) + T_{2k}(v))}_{\geq -q_3\phi_r(v)} d\rho_t(v) + \int_{K_{1k}^c \cap \Omega_r} \underbrace{(T_{1k}(v) + T_{2k}(v))}_{\geq -(q_1+q_2)\phi_r(v)} d\rho_t(v) \right) \\ &\geq -d \max\{q_1 + q_2, q_3\} \int \phi_r(v) d\rho_t(v) = -q \int \phi_r(v) d\rho_t(v). \end{aligned}$$

An application of Grönwall’s inequality concludes the proof. □

3.5 Proof of Theorem 2

We now have all necessary tools to conclude the global convergence proof.

Proof (Theorem 2). Lemma 1 provides a bound for the time derivative of $\mathcal{V}(\rho_t)$, given by

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\leq - (2\lambda - \sigma^2) \mathcal{V}(\rho_t) + \sqrt{2} (\lambda + \sigma^2) \sqrt{\mathcal{V}(\rho_t)} \|v_\alpha(\rho_t) - v^*\|_2 \\ &\quad + \frac{\sigma^2}{2} \|v_\alpha(\rho_t) - v^*\|_2^2. \end{aligned} \tag{17}$$

Now we define the time $T \geq 0$ by

$$T := \sup \{t \geq 0 : \mathcal{V}(\rho_{t'}) > \varepsilon \text{ and } \|v_\alpha(\rho_{t'}) - v^*\|_2 < C(t') \text{ for all } t' \in [0, t]\}, \tag{18}$$

where

$$C(t) := \min \left\{ \frac{\tau}{2} \frac{(2\lambda - \sigma^2)}{\sqrt{2}(\lambda + \sigma^2)}, \sqrt{\tau \frac{(2\lambda - \sigma^2)}{\sigma^2}} \right\} \sqrt{\mathcal{V}(\rho_t)}.$$

Combining (17) with (18), we have by construction for all $t \in (0, T)$

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -(1 - \tau) (2\lambda - \sigma^2) \mathcal{V}(\rho_t).$$

Grönwall’s inequality implies the upper bound

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0) \exp \left(-(1 - \tau) (2\lambda - \sigma^2) t \right), \quad \text{for } t \in [0, T]. \tag{19}$$

Accordingly, we note that $\mathcal{V}(\rho_t)$ is a decreasing function in t , which implies the decay of the auxiliary function $C(t)$ as well. Hence, we may bound

$$\max_{t \in [0, T]} \|v_\alpha(\rho_t) - v^*\|_2 \leq \max_{t \in [0, T]} C(t) \leq C(0), \tag{20}$$

$$\max_{t \in [0, T]} \int \|v - v^*\|_2 d\rho_t(v) \leq \max_{t \in [0, T]} \sqrt{2\mathcal{V}(\rho_t)} \leq \sqrt{2\mathcal{V}(\rho_0)}. \tag{21}$$

To conclude that $\mathcal{V}(\rho_T) \leq \varepsilon$, it now remains to check three different cases.

Case $T \geq T^*$: If $T \geq T^*$, we can use the definition of T^* in (8) and the time-evolution bound of $\mathcal{V}(\rho_t)$ in (19) to conclude that $\mathcal{V}(\rho_{T^*}) \leq \varepsilon$. Hence, by definition of T in (18), we find $\mathcal{V}(\rho_T) = \varepsilon$ and $T = T^*$.

Case $T < T^*$ and $\mathcal{V}(\rho_T) = \varepsilon$: Nothing needs to be discussed in this case.

Case $T < T^*$, $\mathcal{V}(\rho_T) > \varepsilon$, and $\|v_\alpha(\rho_T) - v^*\|_2 \geq C(T)$: We shall show that there exists $\alpha_0 > 0$ so that for any $\alpha \geq \alpha_0$ we have $\|v_\alpha(\rho_T) - v^*\|_2 < C(T)$, a contradiction, which proves that the case never occurs. To do so, we define

$$q := \min \left\{ (\eta C(T) / (2\sqrt{d}))^{1/\nu}, \mathcal{E}_\infty \right\} / 2 \quad \text{and} \quad r := \max_{s \in [0, R_0]} \left\{ \max_{v \in B_s^\infty(v^*)} \mathcal{E}(v) \leq q \right\}.$$

By construction, $r \leq R_0$ and $q + \mathcal{E}_r = q + \sup_{v \in B_r^\infty(v^*)} \mathcal{E}(v) \leq 2q \leq \mathcal{E}_\infty$. Furthermore, we note that $q > 0$ since $C(T) > 0$. By continuity of \mathcal{E} there exists $s_q > 0$ such that $\mathcal{E}(v) \leq q$ for all $v \in B_{s_q}^\infty(v^*)$, thus yielding also $r > 0$. Therefore, we can apply Proposition 1 with q and r as above together with (21) to get

$$\|v_\alpha(\rho_T) - v^*\|_2 \leq \frac{1}{2} C(T) + \frac{\sqrt{d} \exp(-\alpha q)}{\rho_T(B_r^\infty(v^*))} \sqrt{2\mathcal{V}(\rho_0)}. \tag{22}$$

Furthermore, by (20) we have $\max_{t \in [0, T]} \|v_\alpha(\rho_t) - v^*\|_2 \leq C(0)$, implying that all assumptions of Proposition 2 hold. Therefore, there exists $a > 0$ so that

$$\rho_T(B_r^\infty(v^*)) \geq \int \phi_r(v) d\rho_0(v) \exp(-aT) > 0,$$

where we used (7) for bounding the initial mass ρ_0 , and the fact that ϕ_r is bounded from below on $B_{cr}^\infty(v^*)$ for any $c < 1$. Then, by using any $\alpha > \alpha_0$ with

$$\alpha_0 = \frac{1}{2q} \left(\log d - 2 \log \rho_T(B_r^\infty(v^*)) + \log \left(\frac{\mathcal{V}(\rho_0)}{\mathcal{V}(\rho_T)} \right) + 2 \log \left(\frac{\lambda + \sigma^2}{\tau (2\lambda - \sigma^2)} \right) \right),$$

(22) is strictly smaller than $C(T)$, giving the desired contradiction. □

4 A Machine Learning Example

In this section, we showcase the practicability of the implementation of anisotropic CBO as described in [4, Algorithm 2.1] for problems appearing in machine learning by training a shallow and a convolutional NN (CNN) classifier for the MNIST dataset of handwritten digits [16]. Let us emphasize that it is not our aim to challenge the state of the art for this task by employing the most sophisticated model or intricate data preprocessing. We merely believe that this is a well-understood, complex, high-dimensional benchmark to demonstrate that CBO achieves good results already with limited computational capacities.

Let us now describe the NN architectures used in our numerical experiment, see also Fig. 3. Each input image is represented by a matrix of dimension 28×28 with entries valued between 0 and 1 depending on the grayscale of the respective pixel. For the shallow neural net (see Fig. 3a) the image is first reshaped to a vector $x \in \mathbb{R}^{728}$ before being passed through a dense layer of the form $\text{ReLU}(Wx+b)$ with trainable weight matrix $W \in \mathbb{R}^{10 \times 728}$ and bias vector $b \in \mathbb{R}^{10}$. The CNN (see Fig. 3b) has learnable kernels and its architecture is similar to the one of the LeNet-1, cf. [15, Section III.C.7]. In both networks a batch normalization step is included after each ReLU activation, which entails a considerably faster training process. Moreover, in the final layers a softmax activation function is applied so that the output can be interpreted as a probability distribution over the digits. In total, the number of unknowns to be trained in case of the shallow NN is 7850, which compares to 2112 free parameters for the CNN. We denote the parameters of the NN by θ and its forward pass by f_θ .

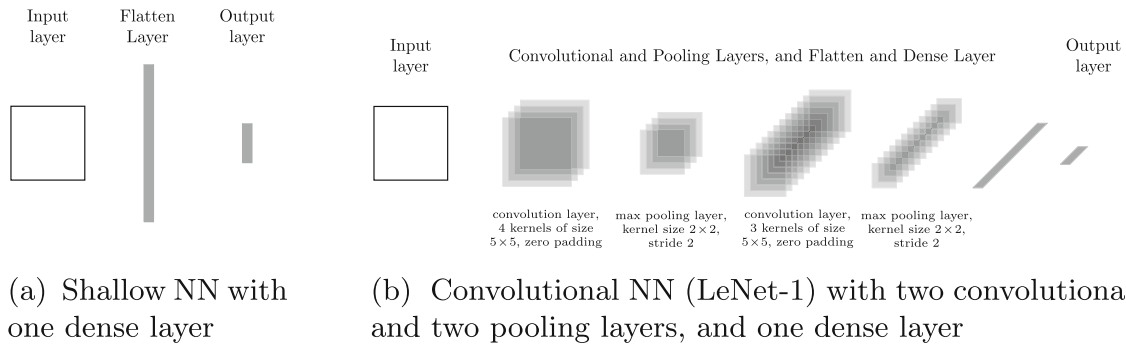


Fig. 3. Architectures of the NNs used in the experiments of Sect. 4.

As a loss function during training we use the categorical crossentropy loss $\ell(\hat{y}, y) = -\sum_{k=0}^9 y_k \log(\hat{y}_k)$ with $\hat{y} = f_\theta(x)$ denoting the output of the NN for a training sample (x, y) consisting of image and label. This gives rise to the objective function $\mathcal{E}(\theta) = \frac{1}{M} \sum_{m=1}^M \ell(f_\theta(x^m), y^m)$, where $(x^m, y^m)_{m=1}^M$ denote the M training samples. When evaluating the performance of the NN we determine the accuracy on a test set by counting the number of successful predictions.

The used implementation of anisotropic CBO combines ideas presented in [10, Section 2.2] with the algorithm proposed in [4]. More precisely, it employs random

mini-batch ideas when evaluating the objective function \mathcal{E} and when computing the consensus point v_α , meaning that \mathcal{E} is only evaluated on a random subset of size $n_\mathcal{E}$ of the training dataset and v_α is only computed from a random subset of size n_N of all particles. While this reduces the computational complexity, it simultaneously increases the stochasticity, which enhances the ability to escape from local optima. Furthermore, inspired by Simulated Annealing, a cooling strategy for the parameters α and σ is used as well as a variance-based particle reduction technique similar to ideas from Genetic Algorithms. More specifically, α is multiplied by 2 after each epoch, while the diffusion parameter σ follows the schedule $\sigma_{epoch} = \sigma_0 / \log_2(epoch + 2)$. For our experiments we choose the parameters $\lambda = 1$, $\sigma_0 = \sqrt{0.4}$ and $\alpha_{initial} = 50$, and discrete time step size $\Delta t = 0.1$ for training both the shallow and the convolutional NN. We use $N = 100$ particles, which are initialized according to $\mathcal{N}((0, \dots, 0)^T, \text{Id})$. The mini-batch sizes are $n_\mathcal{E} = 60$ and $n_N = 10$ and despite v_α being computed only on a basis of n_N particles, all N particles are updated in every time step, referred to as the full update in [4]. We emphasize that hyperparameters have not been tuned extensively.

In Fig. 4 we report the results of our experiment. While achieving a test accuracy of almost 90% for the shallow NN, we obtain around 97% accuracy with the CNN. For comparison, when trained with backpropagation with finely tuned parameters, a comparable CNN achieves 98.3% accuracy, cf. [15, Figure 9]. In view of these results, CBO can be regarded as a successful optimizer for machine learning tasks, which performs comparably to the state of the art. At the same time it is worth highlighting that CBO is extremely versatile and customizable, does not require gradient information or substantial hyperparameter tuning and has the potential to be parallelized.

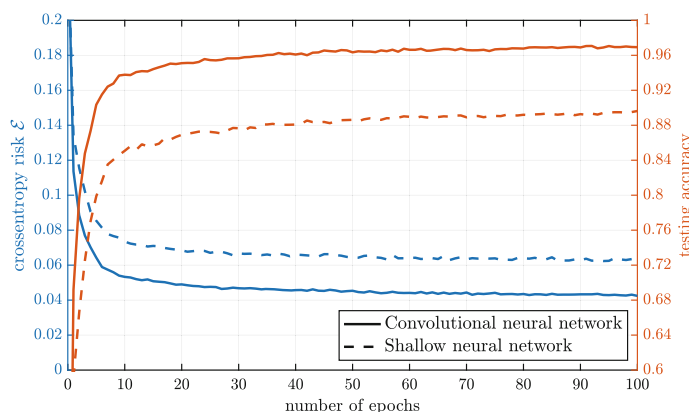


Fig. 4. Comparison of the performances of a shallow (dashed lines) and convolutional (solid lines) NN with architectures as described in Figs. 3a and b, when trained with a discretized version of the anisotropic CBO dynamics (1). Depicted are the accuracies on a test dataset (orange lines) and the values of the objective function \mathcal{E} (blue lines), which was chosen to be the categorical crossentropy loss on a random sample of the training set of size 10000. (Color figure online)

5 Conclusion

In this paper we establish the global convergence of anisotropic consensus-based optimization (CBO) to the global minimizer in mean-field law with dimension-independent convergence rate by adapting the proof technique developed in [9]. It is based on the insight that the dynamics of individual particles follow, on average, the gradient flow dynamics of the map $v \mapsto \|v - v^*\|_2^2$. Furthermore, by utilizing the implementation of anisotropic CBO suggested in [4], we demonstrate the practicability of the method by training the well-known LeNet-1 on the MNIST data set, achieving around 97% accuracy after few epochs with just 100 particles.

In subsequent work we plan to extend our theoretical understanding of CBO to the finite particle regime, and aim to provide extensive numerical studies. We also intend to use this approach to explain the mean-field law convergence behavior of other metaheuristics such as Particle Swarm Optimization, see, e.g., [5, 13].

Acknowledgements. MF acknowledges the support of the DFG Project “Identification of Energies from Observations of Evolutions” and the DFG SPP 1962 “Non-smooth and Complementarity-Based Distributed Parameter Systems: Simulation and Hierarchical Optimization”. KR acknowledges the financial support from the Technical University of Munich – Institute for Ethics in Artificial Intelligence (IEAI).

References

1. Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines. A Stochastic Approach to Combinatorial Optimization and Neural Computing. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, Hoboken (1989)
2. Carrillo, J.A., Choi, Y.P., Totzeck, C., Tse, O.: An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.* **28**(6), 1037–1066 (2018)
3. Carrillo, J.A., Fornasier, M., Toscani, G., Vecil, F.: Particle, kinetic, and hydrodynamic models of swarming. In: *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, pp. 297–336. *Model. Simul. Sci. Eng. Technol.*, Birkhäuser Boston, Boston (2010)
4. Carrillo, J.A., Jin, S., Li, L., Zhu, Y.: A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.* **27**(suppl.), Paper No. S5, 22 (2021)
5. Cipriani, C., Huang, H., Qiu, J.: Zero-inertia limit: from particle swarm optimization to consensus based optimization. [arXiv:2104.06939](https://arxiv.org/abs/2104.06939) (2021)
6. Dorigo, M., Blum, C.: Ant colony optimization theory: a survey. *Theor. Comput. Sci.* **344**(2–3), 243–278 (2005)
7. Fogel, D.B.: *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, 2nd edn. IEEE Press, Piscataway (2000). <https://ieeexplore.ieee.org/book/5237910>
8. Fornasier, M., Huang, H., Pareschi, L., Sünnen, P.: Anisotropic diffusion in consensus-based optimization on the sphere. [arXiv:2104.00420](https://arxiv.org/abs/2104.00420) (2021)

9. Fornasier, M., Klock, T., Riedl, K.: Consensus-based optimization methods converge globally in mean-field law. [arXiv:2103.15130](#) (2021)
10. Fornasier, M., Pareschi, L., Huang, H., Sünnen, P.: Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.* **22**(237), 1–55 (2021)
11. Holland, J.H.: *Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* University of Michigan Press, Ann Arbor (1975)
12. Huang, H., Qiu, J.: On the mean-field limit for the consensus-based optimization. [arXiv:2105.12919](#) (2021)
13. Huang, H., Qiu, J., Riedl, K.: On the global convergence of particle swarm optimization methods. [arXiv:2201.12460](#) (2022)
14. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: *Proceedings of ICNN 1995 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948. IEEE (1995)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
16. LeCun, Y., Cortes, C., Burges, C.: MNIST handwritten digit database (2010). <http://yann.lecun.com/exdb/mnist/>
17. Miller, P.D.: *Applied Asymptotic Analysis.* Graduate Studies in Mathematics, vol. 75. American Mathematical Society, Providence (2006)
18. Pinnau, R., Totzeck, C., Tse, O., Martin, S.: A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.* **27**(1), 183–204 (2017)
19. Rastrigin, L.: The convergence of the random search method in the extremal control of a many parameter system. *Autom. Remote Control* **24**, 1337–1342 (1963)
20. Toscani, G.: Kinetic models of opinion formation. *Commun. Math. Sci.* **4**(3), 481–496 (2006)

License for [CBO-II].

The permission to reprint and include the material is printed on the next page(s).



RightsLink



Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law

SPRINGER NATURE
Author: Massimo Fornasier, Timo Klock, Konstantin Riedl

Publication: Springer eBook

Publisher: Springer Nature

Date: Jan 1, 2022

Copyright © 2022, Springer Nature

Order Completed

Thank you for your order.

This Agreement between Konstantin Riedl ("You") and Springer Nature ("Springer Nature") consists of your order details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License number Reference confirmation email for license number

License date Mar, 28 2024

Licensed Content

Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law
Licensed Content Author	Massimo Fornasier, Timo Klock, Konstantin Riedl
Licensed Content Date	Jan 1, 2022

Order Details

Type of Use	Thesis/Dissertation academic/university or research institute
Requestor type	print and electronic
Format	full article/chapter
Portion	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	yes

About Your Work

Title of new work	Mathematical Foundations of Interacting Multi-Particle Systems for Optimization
Institution name	Technical University of Munich
Expected presentation date	Apr 2024

Additional Data

Requestor Location

Requestor	Mr. Konstantin Riedl
Requestor Location	München, Bayern Germany Attn: Mr. Konstantin Riedl

Tax Details

 **Billing Information**

Billing Type	Invoice Mr. Konstantin Riedl
Billing address	München, Germany Attn: Mr. Konstantin Riedl

Total: 0.00 EUR
CLOSE WINDOW

© 2024 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Data Security and Privacy](#)
| [For California Residents](#) | [Terms and Conditions](#) Comments? We would like to hear from you. E-mail us at customer@copyright.com

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Mar 28, 2024

This Agreement between Mr. Konstantin Riedl ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5757550115464
License date	Mar 28, 2024
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law
Licensed Content Author	Massimo Fornasier, Timo Klock, Konstantin Riedl
Licensed Content Date	Jan 1, 2022
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter

Will you be translating? no

Circulation/distribution 1 - 29

Author of this Springer Nature content yes

Title of new work Mathematical Foundations of Interacting Multi-Particle Systems for Optimization

Institution name Technical University of Munich

Expected presentation date Apr 2024

Mr. Konstantin Riedl

Requestor Location München, Bayern
Germany
Attn: Mr. Konstantin Riedl

Billing Type Invoice

Mr. Konstantin Riedl

Billing Address München, Germany
Attn: Mr. Konstantin Riedl

Total 0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By

clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at <https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/>.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition

Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other

assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United

Nations Convention on Contracts for the International Sale of Goods) or to Germany's choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit <https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution>

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

Paper P3

Consensus-Based Optimization with Truncated Noise

M. Fornasier, P. Richtárik, K. Riedl, and L. Sun

Eur. J. Appl. Math. (special issue “From integro-differential models to data-oriented approaches for emergent phenomena”) (accepted 2024, to appear)

Paper Summary of [CBO-III]³⁶

In the paper “Consensus-Based Optimization with Truncated Noise,” published in the special issue “*From integro-differential models to data-oriented approaches for emergent phenomena*” of the *European Journal of Applied Mathematics*, we propose and explore the variant (3.106) of CBO, which incorporates truncated noise in order to enhance the well-behavedness of statistics of the law of the dynamics.

CBO is a versatile multi-particle metaheuristic optimization method suitable for performing nonconvex and nonsmooth global optimizations in form of (2.1) and comes with an analytical framework [CBO-I; CBO-II], which is flexible enough to allow for various modifications of the dynamics.

By introducing a truncation of the noise term of the CBO dynamics as in (3.106), we achieve in [CBO-III] that, in contrast to the original version of the algorithm, the law of the mean-field dynamics exhibits sub-Gaussian behavior [CBO-III, Lemma 8]. This permits that higher moments of the law of the particle system can be effectively bounded [CBO-III, Section 1]. As a result, our proposed variant exhibits enhanced convergence performance, allowing in particular for wider flexibility in choosing the noise parameter of the method [CBO-III, Theorem 3], which is also confirmed experimentally [CBO-III, Figure 1 and Section 4]. By adopting the analytical framework of [CBO-I] we further rigorously prove global convergence in expectation of the proposed CBO variant requiring only minimal assumptions on the objective function and on the initialization [CBO-III, Theorem 3]. Theoretical improvements with respect to previous works and as a consequence of the sub-Gaussian behavior resulting from the truncated noise are reflected in [CBO-III, Proposition 15], where global convergence in mean-field law is assured under weaker conditions on the noise parameter of the method, as well as in [CBO-III, Proposition 7], where a non-probabilistic quantitative mean-field approximation result is provided. The combination of the former results about the convergence in mean-field law and the quantitative mean-field approximation together with classical results of numerical approximation of SDEs allows to obtain the aforementioned global convergence guarantees in expectation of the numerical CBO method [CBO-III, Theorem 3]. From convergence in expectation, a convergence in probability result can be derived immediately. Numerical evidences demonstrate the benefit of truncating the noise in CBO [CBO-III, Figure 1 and Section 4].

KR’s Contributions. After MF reached out to PR to discuss about CBO and an initial meeting together with KR, LS proposed the idea of using truncated noise in the CBO dynamics and worked out the technical details of the analysis following the analytical framework established by MF, KR, and collaborators in earlier works. LS and KR conducted the numerical experiments. During LS’s visit at the Technical University of Munich, LS and KR discussed, prepared, and finalized the manuscript, which was then proofread by and refined together with MF and PR.

³⁶In this section, we follow [CBO-III, Abstract].

The following document is a reprint of

- [CBO-III] M. Fornasier, P. Richtárik, K. Riedl, and L. Sun. “Consensus-Based Optimization with Truncated Noise.” In: *Eur. J. Appl. Math. (special issue “From integro-differential models to data-oriented approaches for emergent phenomena”)* (accepted 2024, to appear).

The permission to reprint and include the material is provided after the reprint.

Consensus-Based Optimization with Truncated Noise

Massimo Fornasier^{*1,2,3}, Peter Richtárik^{†4,5,6}, Konstantin Riedl^{‡1,2} and Lukang Sun^{§4,5}

¹*Technical University of Munich, School of Computation, Information and Technology,
Department of Mathematics, Munich, Germany*

²*Munich Center for Machine Learning, Munich, Germany*

³*Munich Data Science Institute, Germany*

⁴*King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*

⁵*KAUST AI Initiative, Thuwal, Saudi Arabia*

⁶*SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence,
Thuwal, Saudi Arabia*

Abstract

Consensus-based optimization (CBO) is a versatile multi-particle metaheuristic optimization method suitable for performing nonconvex and nonsmooth global optimizations in high dimensions. It has proven effective in various applications while at the same time being amenable to a theoretical convergence analysis. In this paper, we explore a variant of CBO, which incorporates truncated noise in order to enhance the well-behavedness of the statistics of the law of the dynamics. By introducing this additional truncation in the noise term of the CBO dynamics, we achieve that, in contrast to the original version, higher moments of the law of the particle system can be effectively bounded. As a result, our proposed variant exhibits enhanced convergence performance, allowing in particular for wider flexibility in choosing the noise parameter of the method as we confirm experimentally. By analyzing the time-evolution of the Wasserstein-2 distance between the empirical measure of the interacting particle system and the global minimizer of the objective function, we rigorously prove convergence in expectation of the proposed CBO variant requiring only minimal assumptions on the objective function and on the initialization. Numerical evidences demonstrate the benefit of truncating the noise in CBO.

Keywords: global optimization, derivative-free optimization, nonsmoothness, nonconvexity, metaheuristics, consensus-based optimization, truncated noise

AMS subject classifications: 65K10, 90C26, 90C56, 35Q90, 35Q84

1 Introduction

The search for a global minimizer v^* of a potentially nonconvex and nonsmooth cost function

$$f : \mathbb{R}^d \rightarrow \mathbb{R}$$

*Email: massimo.fornasier@cit.tum.de (corresponding author)

†Email: peter.richtarik@kaust.edu.sa

‡Email: konstantin.riedl@ma.tum.de

§Email: lukang.sun@kaust.edu.sa

holds significant importance in a variety of applications throughout applied mathematics, science and technology, engineering, and machine learning. Historically, a class of methods known as metaheuristics [3,5] has been developed to address this inherently challenging and, in general, NP-hard problem. Examples of such include evolutionary programming [19], genetic algorithms [31], particle swarm optimization (PSO) [36], simulated annealing [1], and many others. These methods work combining local improvement procedures and global strategies by orchestrating deterministic and stochastic advances, with the aim of creating a method capable of robustly and efficiently finding the globally minimizing argument v^* of f . However, despite their empirical success and widespread adoption in practice, most metaheuristics lack a solid mathematical foundation that could guarantee their robust convergence to global minimizers under reasonable assumptions.

Motivated by the urge to devise algorithms which converge provably, a novel class of metaheuristics, so-called consensus-based optimization (CBO), originally proposed by the authors of [40], has recently emerged in the literature. Due to the inherent simplicity in the design of CBO, this class of optimization algorithms lends itself to a rigorous theoretical analysis, as demonstrated in particular in the works [11, 13, 23, 24, 27, 28, 39]. However, this recent line of research does not just offer a promising avenue for establishing a thorough mathematical framework for understanding the numerically observed successes of CBO methods [13, 15, 21, 24, 42], but beyond that allows to explain the effective use of conceptually similar and wide-spread methods such as PSO as well as at first glance completely different optimization algorithms such as stochastic gradient descent (SGD). While the first connection is to be expected and by now made fairly rigorous [17, 26, 34] due to CBO indisputably taking PSO as inspiration, the second observation is somewhat surprising, as it builds a bridge between derivative-free metaheuristics and gradient-based learning algorithms. Despite CBO solely relying on evaluations of the objective function, recent work [43] reveals an intrinsic SGD-like behavior of CBO itself by interpreting it as a certain stochastic relaxation of gradient descent, which provably overcomes energy barriers of nonconvex function. These perspectives, and, in particular the already well-investigated convergence behavior of standard CBO, encourage the exploration of improvements to the method in order to allow overcoming the limitations of traditional metaheuristics mentioned at the start. For recent surveys on CBO we refer to [25, 45].

While the original CBO model [40] has been adapted to solve constrained optimizations [4, 9, 14], optimizations on manifolds [20–22, 29, 37], multi-objective optimization problems [7, 8, 38], saddle point problems [33] or the task of sampling [12], as well as has been extended to make use of memory mechanisms [6, 42, 46], gradient information [42, 44], momentum [16], jump-diffusion processes [35] or localization kernels for polarization [10], we focus in this work on a variation of the original model, which incorporates a truncation in the noise term of the dynamics. More formally, given a time horizon $T > 0$, a time discretization $t_0 = 0 < \Delta t < \dots < K\Delta t = t_K = T$ of $[0, T]$, and user-specified parameters $\alpha, \lambda, \sigma > 0$ as well as $v_b, R > 0$, we consider the interacting particle system

$$V_{k+1, \Delta t}^i - V_{k, \Delta t}^i = -\Delta t \lambda \left(V_{k, \Delta t}^i - \mathcal{P}_{v_b, R} \left(v_\alpha(\widehat{\rho}_{k, \Delta t}^N) \right) \right) + \sigma \left(\|V_{k, \Delta t}^i - v_\alpha(\widehat{\rho}_{k, \Delta t}^N)\|_2 \wedge M \right) B_{k, \Delta t}^i, \quad (1)$$

$$V_0^i \sim \rho_0 \quad \text{for all } i = 1, \dots, N, \quad (2)$$

where $((B_{k, \Delta t}^i)_{k=0, \dots, K-1})_{i=1, \dots, N}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}_d$. Equation (1) originates from a simple Euler-Maruyama time discretization [30, 41] of the system of stochastic differential

equations (SDEs), expressed in Itô's form as

$$dV_t^i = -\lambda \left(V_t^i - \mathcal{P}_{v_b, R} \left(v_\alpha(\widehat{\rho}_t^N) \right) \right) dt + \sigma \left(\|V_t^i - v_\alpha(\widehat{\rho}_t^N)\|_2 \wedge M \right) dB_t^i \quad (3)$$

$$V_0^i \sim \rho_0 \quad \text{for all } i = 1, \dots, N. \quad (4)$$

where $((B_t^i)_{t \geq 0})_{i=1, \dots, N}$ are now independent standard Brownian motions in \mathbb{R}^d . The empirical measure of the particles at time t is denoted by $\widehat{\rho}_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{V_t^i}$. Moreover, $\mathcal{P}_{v_b, R}$ is the projection onto $B_R(v_b)$ defined as

$$\mathcal{P}_{v_b, R}(v) := \begin{cases} v, & \text{if } \|v - v_b\|_2 \leq R, \\ v_b + R \frac{v - v_b}{\|v - v_b\|_2}, & \text{if } \|v - v_b\|_2 > R. \end{cases} \quad (5)$$

As a crucial assumption in this paper, the map $\mathcal{P}_{v_b, R}$ depends on R and v_b in such way that $v^* \in B_R(v_b)$. Setting the parameters can be feasible under specific circumstances, as exemplified by the regularized optimization problem $f(v) := \text{Loss}(v) + \Lambda \|v\|_2$, wherein $v^* \in B_{\text{Loss}(0)/\Lambda}(0)$. In the absence of prior knowledge regarding v_b and R , a practical approach is to choose $v_b = 0$ and assign a sufficiently large value to R . The first terms in (1) and (3), respectively, impose a deterministic drift of each particle towards the possibly projected momentaneous consensus point $v_\alpha(\widehat{\rho}_t^N)$, which is a weighted average of the particles' positions and computed according to

$$v_\alpha(\widehat{\rho}_t^N) := \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L^1(\widehat{\rho}_t^N)}} d\widehat{\rho}_t^N(v). \quad (6)$$

The weights $\omega_\alpha(v) := \exp(-\alpha f(v))$ are motivated by the well-known Laplace principle [18], which states for any absolutely continuous probability distribution ϱ on \mathbb{R}^d that

$$\lim_{\alpha \rightarrow \infty} \left(-\frac{1}{\alpha} \log \left(\int \omega_\alpha(v) d\varrho(v) \right) \right) = \inf_{v \in \text{supp}(\varrho)} f(v) \quad (7)$$

and thus justifies that $v_\alpha(\widehat{\rho}_t^N)$ serves as a suitable proxy for the global minimizer v^* given the currently available information of the particles $(V_t^i)_{i=1, \dots, N}$. The second terms in (1) and (3), respectively, encode the diffusion or exploration mechanism of the algorithm, where, in contrast to standard CBO, we truncate the noise by some fixed constant $M > 0$.

We conclude and re-iterate that both the introduction of the projection $\mathcal{P}_{v_b, R}(v_\alpha(\widehat{\rho}_t^N))$ of the consensus point and the employment of truncation of the noise variance $(\|V_t^i - v_\alpha(\widehat{\rho}_t^N)\|_2 \wedge M)$ are main innovations to the original CBO method. We shall explain and justify these modifications in the following paragraph.

Despite these technical improvements, the approach to analyze the convergence behavior of the implementable scheme (1) follows a similar route already explored in [11, 13, 23, 24]. In particular, the convergence behavior of the method to the global minimizer v^* of the objective f is investigated on the level of the mean-field limit [23, 32] of the system (3). More precisely, we study the macroscopic behavior of the agent density $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, where $\rho_t = \text{Law}(\bar{V}_t)$ with

$$d\bar{V}_t = -\lambda \left(\bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) \right) dt + \sigma \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2 \wedge M \right) dB_t \quad (8)$$

and initial data $\bar{V}_0 \sim \rho_0$. Afterwards, by establishing a quantitative estimate on the mean-field approximation, i.e., the proximity of the mean-field system (8) to the interacting particle system (3) and combining the two results, we obtain a convergence result for the CBO algorithm (1) with truncated noise.

Motivation for using truncated noise. In what follows we provide a heuristic explanation of the theoretical benefits of employing a truncation in the noise of CBO as in (1), (3) and (8). Let us therefore first recall that the standard variant of CBO [40] can be retrieved from the model considered in this paper by setting $v_b = 0$, $R = \infty$ and $M = \infty$. For instance, in place of the mean-field dynamics (8), we would have

$$d\bar{V}_t^{\text{CBO}} = -\lambda \left(\bar{V}_t^{\text{CBO}} - v_\alpha(\rho_t^{\text{CBO}}) \right) dt + \sigma \left\| \bar{V}_t^{\text{CBO}} - v_\alpha(\rho_t^{\text{CBO}}) \right\|_2 dB_t.$$

Attributed to the Laplace principle (7) it holds $v_\alpha(\rho_t^{\text{CBO}}) \approx v^*$ for α sufficiently large, i.e., as $\alpha \rightarrow \infty$, the former dynamics converges to

$$d\bar{Y}_t^{\text{CBO}} = -\lambda \left(\bar{Y}_t^{\text{CBO}} - v^* \right) dt + \sigma \left\| \bar{Y}_t^{\text{CBO}} - v^* \right\|_2 dB_t. \quad (9)$$

Firstly, observe that here the first term imposes a direct drift to the global minimizer v^* and thereby induces a contracting behavior, which is on the other hand counteracted by the diffusion term, which contributes a stochastic exploration around this point. In particular, with \bar{Y}_t^{CBO} approaching v^* , the exploration vanishes so that \bar{Y}_t^{CBO} converges eventually deterministically to v^* . Conversely, as long as \bar{Y}_t^{CBO} is far away from v^* , the order of the random exploration is strong. By Itô's formula we have

$$\frac{d}{dt} \mathbb{E} \left[\left\| \bar{Y}_t^{\text{CBO}} - v^* \right\|_2^p \right] = p \left(-\lambda + \frac{\sigma^2}{2} (p + d - 2) \right) \mathbb{E} \left[\left\| \bar{Y}_t^{\text{CBO}} - v^* \right\|_2^p \right]$$

and thus

$$\mathbb{E} \left[\left\| \bar{Y}_t^{\text{CBO}} - v^* \right\|_2^p \right] = \exp \left(p \left(-\lambda + \frac{\sigma^2}{2} (p + d - 2) \right) t \right) \mathbb{E} \left[\left\| \bar{Y}_0^{\text{CBO}} - v^* \right\|_2^p \right] \quad (10)$$

for any $p \geq 1$. Denoting with μ_t^{CBO} the law of \bar{Y}_t^{CBO} , this means that, given any $\lambda, \sigma > 0$, there is some threshold exponent $p^* = p^*(\lambda, \sigma, d)$, such that

$$\begin{aligned} \lim_{t \rightarrow \infty} W_p \left(\mu_t^{\text{CBO}}, \delta_{v^*} \right) &= \lim_{t \rightarrow \infty} \left(\mathbb{E} \left[\left\| \bar{Y}_t^{\text{CBO}} - v^* \right\|_2^p \right] \right)^{1/p} \\ &= \lim_{t \rightarrow \infty} \exp \left(\left(-\lambda + \frac{\sigma^2}{2} (p + d - 2) \right) t \right) \left(\mathbb{E} \left[\left\| \bar{Y}_0^{\text{CBO}} - v^* \right\|_2^p \right] \right)^{1/p} \\ &= 0 \end{aligned}$$

for $p < p^*$, while for $p > p^*$ it holds

$$\begin{aligned} \lim_{t \rightarrow \infty} W_p \left(\mu_t^{\text{CBO}}, \delta_{v^*} \right) &= \lim_{t \rightarrow \infty} \left(\mathbb{E} \left[\left\| \bar{Y}_t^{\text{CBO}} - v^* \right\|_2^p \right] \right)^{1/p} \\ &= \lim_{t \rightarrow \infty} \exp \left(\left(-\lambda + \frac{\sigma^2}{2} (p + d - 2) \right) t \right) \left(\mathbb{E} \left[\left\| \bar{Y}_0^{\text{CBO}} - v^* \right\|_2^p \right] \right)^{1/p} \\ &= \infty. \end{aligned}$$

Recalling that the distribution of a random variable Y has heavy tails if and only if the moment generating function $M_Y(s) := \mathbb{E} [\exp(sY)] = \mathbb{E} \left[\sum_{p=0}^{\infty} (sY)^p / p! \right]$ is infinite for all $s > 0$, these computations suggest that the distribution of μ_t^{CBO} exhibits characteristics of heavy tails as $t \rightarrow \infty$, thereby increasing the likelihood of encountering outliers in a sample drawn from μ_t^{CBO} for large t .

On the contrary, for CBO with truncated noise (8), we get, thanks once again to the Laplace principle as $\alpha \rightarrow \infty$, that (8) converges to

$$d\bar{Y}_t = -\lambda \left(\bar{Y}_t - v^* \right) dt + \sigma \left(\left\| \bar{Y}_t - v^* \right\|_2 \wedge M \right) dB_t, \quad (11)$$

for which we can compute

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{Y}_t - v^*\|_2^p \right] &\leq -p\lambda \mathbb{E} \left[\|\bar{Y}_t - v^*\|_2^p \right] + p \frac{\sigma^2}{2} M^2 (p + d - 2) \mathbb{E} \left[\|\bar{Y}_t - v^*\|_2^{p-2} \right] \\ &\leq -\lambda \mathbb{E} \left[\|\bar{Y}_t - v^*\|_2^p \right] + \lambda \frac{\sigma^p M^p (d + p - 2)^{\frac{p}{2}}}{\lambda^{\frac{p}{2}}}, \end{aligned}$$

for any $p \geq 2$. Notice, that to obtain the second inequality we used Young's inequality¹ as well as Jensen's inequality. By means of Grönwall's inequality, we then have

$$\mathbb{E} \left[\|\bar{Y}_t - v^*\|_2^p \right] \leq \exp(-\lambda t) \mathbb{E} \left[\|\bar{Y}_0 - v^*\|_2^p \right] + \frac{\sigma^p M^p (d + p - 2)^{\frac{p}{2}}}{\lambda^{\frac{p}{2}}} \quad (12)$$

and therefore, denoting with μ_t the law of \bar{Y}_t ,

$$\lim_{t \rightarrow \infty} W_p(\mu_t, \delta_{v^*}) \leq \frac{\sigma M \sqrt{d + p - 2}}{\lambda^{\frac{1}{2}}} < \infty$$

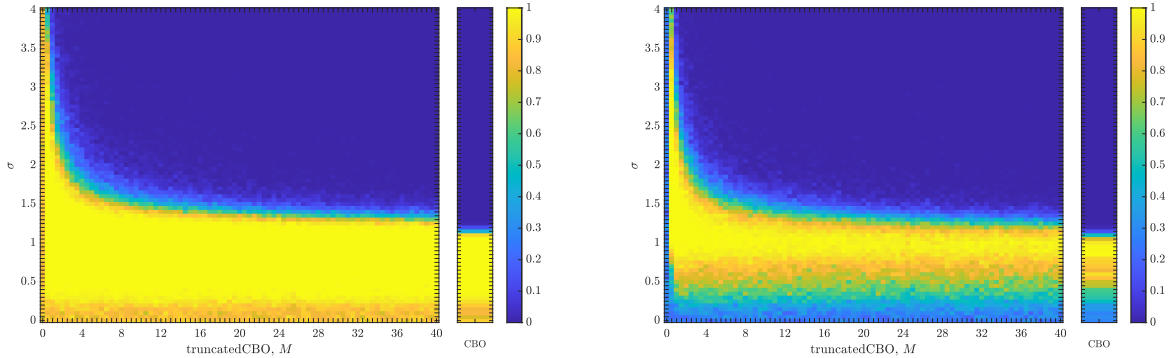
for any $p \geq 2$.

In conclusion, we observe from Equation (10) that the standard CBO dynamics as described in Equation (9) diverges in the setting $\sigma^2 d > 2\lambda$ when considering the Wasserstein-2 distance W_2 . Contrarily, according to Equation (12), the CBO dynamics with truncated noise as presented in Equation (11) converges with exponential rate towards a neighborhood of v^* , with radius $\sigma M \sqrt{d} / \sqrt{\lambda}$. This implies that for a relatively small value of M the CBO dynamics with truncated noise exhibits greater robustness in relation to the parameter $\sigma^2 d / \lambda$. This effect is confirmed numerically in Figure 1.

Remark 1 (Sub-Gaussianity of truncated CBO). *An application of Itô's formula allows to show that, for some $\kappa > 0$, $\mathbb{E} \left[\exp \left(\|\bar{Y}_t - v^*\|_2^2 / \kappa^2 \right) \right] < \infty$, provided $\mathbb{E} \left[\exp \left(\|\bar{Y}_0 - v^*\|_2^2 / \kappa^2 \right) \right] < \infty$. Thus, by incorporating a truncation in the noise term of the CBO dynamics, we ensure that the resulting distribution μ_t exhibits sub-Gaussian behavior and therefore we enhance the regularity and well-behavedness of the statistics of μ_t . As a consequence, more reliable and stable results when analyzing the properties and characteristics of the dynamics are to be expected.*

Contributions. In view of the aforementioned enhanced regularity and well-behavedness of the statistics of CBO with truncated noise compared to standard CBO [40] together with the numerically observed improved performance as depicted in Figure 1, a rigorous convergence analysis of the implementable CBO algorithm with truncated noise as given in (1) is of theoretical interest. In this work we provide theoretical guarantees of global convergence of (1) to the global minimizer v^* for possibly nonconvex and nonsmooth objective functions f . The approach to analyze the convergence behavior of the implementable scheme (1) follows a similar route as initiated and explored by the authors of [11, 13, 23, 24]. In particular, we first investigate the mean-field behavior (8) of the system (3). Then, by establishing a quantitative estimate on the mean-field approximation, i.e., the proximity of the mean-field system (8) to the interacting particle system (3), we obtain a convergence result for the CBO algorithm (1) with truncated noise. Our proving technique nevertheless differs in crucial parts from the one in [23, 24] as, on the one side, we do take advantage of the truncations, and, on the other side, we require additional technical effort to exploit and deal with the enhanced flexibility of the truncated model. Specifically, the central novelty can be identified in the proof of sub-Gaussianity of the process, see Lemma 8.

¹Choose $a = \lambda^{\frac{p-2}{p}} \mathbb{E} \left[\|\bar{Y}_t - v^*\|_2^{p-2} \right]$ and $b = \frac{\sigma^2 M^2 (d+p-2)}{\lambda^{(p-2)/p}}$, and recall that $ab \leq \frac{p-2}{p} a \frac{p}{p-2} + \frac{2}{p} b^{\frac{p}{2}}$.



(a) Phase diagram of success probabilities of isotropic CBO with and without truncated noise at the example of the Ackley function $f(v) = -20 \exp(-0.2/\sqrt{d} \|v\|_2) - \exp(1/d \sum_{k=1}^d \cos(2\pi v_k))$ with $d = 4$

(b) Phase diagram of success probabilities of isotropic CBO with and without truncated noise at the example of the Rastrigin function $f(v) = \sum_{k=1}^d v_k^2 + 2.5(1 - \cos(2\pi v_k))$ with $d = 4$

Figure 1: A comparison of the success probabilities of isotropic CBO with (left phase diagrams) and without (right separate columns) truncated noise for different values of the truncation parameter M and the noise level σ . (Note that standard CBO as investigated in [11, 23, 40] is retrieved when choosing $M = \infty$, $R = \infty$ and $v_b = 0$ in (1)). In both settings (a) and (b) the depicted success probabilities are averaged over 100 runs and the implemented scheme is given by an Euler-Maruyama discretization of Equation (3) with time horizon $T = 50$, discrete time step size $\Delta t = 0.01$, $R = \infty$, $v_b = 0$, $\alpha = 10^5$ and $\lambda = 1$. We use $N = 100$ particles, which are initialized according to $\rho_0 = \mathcal{N}((1, \dots, 1), 2000)$. In both figures we plot the success probability of standard CBO (right separate column) and the CBO variant with truncated noise (left phase transition diagram) for different values of the truncation parameter M and the noise level σ , when optimizing the Ackley ((a)) and Rastrigin ((b)) function, respectively. We observe that truncating the noise term (by decreasing M) consistently allows for a wider flexibility when choosing the noise level σ and thus increasing the likelihood of successfully locating the global minimizer.

1.1 Organization

In Section 2 we present and discuss our main theoretical contribution about the global convergence of CBO with truncated noise in probability and expectation. Section 3 collects the necessary proof details for this result. In Section 4 we numerically demonstrate the benefits of using truncated noise, before we provide a conclusion of the paper in Section 5. For the sake of reproducible research, in the GitHub repository <https://github.com/KonstantinRiedl/CBOGlobalConvergenceAnalysis> we provide the Matlab code implementing CBO with truncated noise.

1.2 Notation

We use $\|\cdot\|_2$ to denote the Euclidean norm on \mathbb{R}^d . Euclidean balls are denoted as $B_r(u) := \{v \in \mathbb{R}^d : \|v - u\|_2 \leq r\}$. For the space of continuous functions $f : X \rightarrow Y$ we write $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$ and a suitable topological space Y . For an open set $X \subset \mathbb{R}^n$ and for $Y = \mathbb{R}^m$ the spaces $\mathcal{C}_c^k(X, Y)$ and $\mathcal{C}_b^k(X, Y)$ contain functions $f \in \mathcal{C}(X, Y)$ that are k -times continuously differentiable and have compact support or are bounded, respectively. We omit Y in the real-valued case. All stochastic processes are considered on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The main objects of study are laws of such processes, $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, where the set $\mathcal{P}(\mathbb{R}^d)$ contains all Borel probability measures over \mathbb{R}^d . With $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ we refer to a snapshot of such law at time t . Measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite p -th moment $\int \|v\|_2^p d\varrho(v)$ are collected in $\mathcal{P}_p(\mathbb{R}^d)$. For any $1 \leq p < \infty$, W_p denotes the Wasserstein- p distance between two Borel probability measures $\varrho_1, \varrho_2 \in \mathcal{P}_p(\mathbb{R}^d)$, see, e.g., [2]. $\mathbb{E}[\cdot]$ denotes the expectation.

2 Global Convergence of CBO with Truncated Noise

We now present the main theoretical result of this work about the global convergence of CBO with truncated noise for objective functions that satisfy the following conditions.

Definition 2 (Assumptions). *Throughout we are interested in functions $f \in \mathcal{C}(\mathbb{R}^d)$, for which*

A1 there exist $v^ \in \mathbb{R}^d$ such that $f(v^*) = \inf_{v \in \mathbb{R}^d} f(v) =: \underline{f}$ and $\underline{\alpha}, L_u > 0$ such that*

$$\sup_{v \in \mathbb{R}^d} \left\| v e^{-\alpha(f(v) - \underline{f})} \right\|_2 =: L_u < \infty \quad (13)$$

for any $\alpha \geq \underline{\alpha}$ and any $v \in \mathbb{R}^d$,

A2 there exist $f_\infty, R_0, \nu, L_\nu > 0$ such that

$$\|v - v^*\|_2 \leq \frac{1}{L_\nu} (f(v) - \underline{f})^\nu \quad \text{for all } v \in B_{R_0}(v^*), \quad (14)$$

$$f_\infty < f(v) - \underline{f} \quad \text{for all } v \in (B_{R_0}(v^*))^c, \quad (15)$$

A3 there exist $L_\gamma > 0, \gamma \in [0, 1]$ such that

$$|f(v) - f(w)| \leq L_\gamma (\|v - v^*\|_2^\gamma + \|w - v^*\|_2^\gamma) \|v - w\|_2 \quad \text{for all } v, w \in \mathbb{R}^d, \quad (16)$$

$$f(v) - \underline{f} \leq L_\gamma (1 + \|v - v^*\|_2^{1+\gamma}) \quad \text{for all } v \in \mathbb{R}^d. \quad (17)$$

A few comments are in order: Condition A1 establishes the existence of a minimizer v^* and requires a certain growth of the function f . Condition A2 ensures that the value of the function f at a point v can locally be an indicator of the distance between v and the minimizer v^* . This error bound condition was first introduced in [23] under the name inverse continuity condition. It in particular guarantees the uniqueness of the global minimizer v^* . Condition A3 sets controllable bounds on the local Lipschitz constant of f and on the growth of f , which is required to be at most quadratic. A similar requirement appears also in [11, 23], but there also a quadratic lower bound was imposed.

2.1 Main Result

We can now state the main result of the paper. Its proof is deferred to Section 3.

Theorem 3. *Let $f \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1, A2 and A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ with $v^* \in \text{supp}(\rho_0)$. Let $V_{0, \Delta t}^i$ be sampled i.i.d. from ρ_0 and denote by $((V_{k, \Delta t}^i)_{k=1, \dots, K})_{i=1, \dots, N}$ the iterations generated by the numerical scheme (1). Fix any $\epsilon \in (0, W_2^2(\rho_0, \delta_{v^*}))$, define the time horizon*

$$T^* := \frac{1}{\lambda} \log \left(\frac{2W_2^2(\rho_0, \delta_{v^*})}{\epsilon} \right)$$

and let $K \in \mathbb{N}$ and Δt satisfy $K\Delta t = T^$. Moreover, let $R \in (\|v_b - v^*\|_2 + \sqrt{\epsilon/2}, \infty)$, $M \in (0, \infty)$ and $\lambda, \sigma > 0$ be such that $\lambda \geq 2\sigma^2 d$ or $\sigma^2 M^2 d = \mathcal{O}(\epsilon)$. Then, by choosing α sufficiently large and $N \geq (16\alpha L_\gamma \sigma^2 M^2)/\lambda$, it holds*

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N V_{K, \Delta t}^i - v^* \right\|_2^2 \right] \lesssim C_{\text{NA}} (\Delta t)^{2m} + \frac{C_{\text{MFA}}}{N} + \epsilon \quad (18)$$

up to a generic constant. Here, C_{NA} depends linearly on the dimension d and the number of particles N and exponentially on the time horizon T^ , m is the order of accuracy of the numerical scheme (for the Euler-Maruyama scheme $m = 1/2$), and $C_{\text{MFA}} = C_{\text{MFA}}(\lambda, \sigma, d, \alpha, L_\nu, \nu, L_\gamma, L_u, T^*, R, v_b, v^*, M)$.*

Remark 4. In the statement of Theorem 3, the parameters R and v_b play a crucial role. We already mentioned how they can be chosen in an example after Equation (5). The role of these parameters is bolstered in particular in the proof of Theorem 3, where it is demonstrated that, by selecting a sufficiently large α depending on R and v_b , the dynamics (8) can be set equal to

$$d\bar{V}_t = -\lambda (\bar{V}_t - \mathcal{P}_{v^*, \delta}(v_\alpha(\rho_t))) dt + \sigma (\|\bar{V}_t - v_\alpha(\rho_t)\|_2 \wedge M) dB_t,$$

where δ represents a small value. For the dynamics (3), we can analogously establish its equivalence to

$$dV_t^i = -\lambda (V_t^i - \mathcal{P}_{v^*, \delta}(v_\alpha(\hat{\rho}_t^N))) dt + \sigma (\|V_t^i - v_\alpha(\hat{\rho}_t^N)\|_2 \wedge M) dB_t^i, \quad i = 1, \dots, N,$$

with high probability, contingent upon the selection of sufficiently large values for both α and N .

Remark 5. The convergence result in form of Theorem 3 obtained in this work differs from the one presented in [23, Theorem 14] in the sense that we obtain convergence in expectation, while in [23] convergence with high probability is established. This distinction arises from the truncation of the noise term employed in our algorithm.

3 Proof Details for Section 2

3.1 Well-Posedness of Equations (1) and (3)

With the projection map $\mathcal{P}_{v_b, R}$ being 1-Lipschitz, existence and uniqueness of strong solutions to the SDEs (1) and (3) are assured by essentially analogous proofs as in [11, Theorems 2.1, 3.1 and 3.2]. The details shall be omitted. Let us remark, however, that due to the presence of the truncation and the projection map, we do not require the function f to be bounded from above or exhibit quadratic growth outside a ball, as required in [11, Theorems 2.1, 3.1 and 3.2].

3.2 Proof Details for Theorem 3

Remark 6. Since adding some constant offset to f does not affect the dynamics of Equations (3) and (8), we will assume $\underline{f} = 0$ in the proofs for simplicity but without loss of generality.

Let us first provide a sketch of the proof of Theorem 3. For the approximation error (18) we have the error decomposition

$$\begin{aligned} \mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N V_{K, \Delta t}^i - v^* \right\|_2^2 \right] &\lesssim \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (V_{K, \Delta t}^i - V_{T^*}^i) \right\|_2^2 \right]}_I + \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (V_{T^*}^i - \bar{V}_{T^*}^i) \right\|_2^2 \right]}_{II} \\ &\quad + \underbrace{\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_{T^*}^i - v^* \right\|_2^2 \right]}_{III}, \end{aligned} \tag{19}$$

where $((\bar{V}_t^i)_{t \geq 0})_{i=1, \dots, N}$ denote N independent copies of the mean-field process $(\bar{V}_t)_{t \geq 0}$ satisfying Equation (8).

In what follows, we investigate each of the three term separately. Term I can be bounded by $C_{NA} (\Delta t)^{2m}$ using classical results on the convergence of numerical schemes for stochastic differential equations (SDEs), as mentioned for instance in [41]. The second and third term,

respectively, are analyzed in separate subsections, providing detailed explanations and bounds for each of the two terms *II* and *III*.

Before doing so, let us provide a concise guide for reading the proofs. As the proofs are quite technical, we start for reader's convenience by presenting the main building blocks of the result first, and collect the more technical steps in subsequent lemmas. This arrangement should hopefully allow to grasp the structure of the proof more easily, and to dig deeper into the details along with the reading.

3.2.1 Upper Bound for the Second Term in (19)

For Term *II* of the error decomposition (19) we have the following upper bound.

Proposition 7. *Let $f \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1, A2 and A3. Moreover, let R and M be finite such that $R \geq \|v_b - v^*\|_2$ and let $N \geq (16\alpha L_\gamma \sigma^2 M^2)/\lambda$. Then we have*

$$\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N (V_{T^*}^i - \bar{V}_{T^*}^i) \right\|_2^2 \right] \leq \frac{C_{\text{MFA}}}{N}, \quad (20)$$

where $C_{\text{MFA}} = C_{\text{MFA}}(\lambda, \sigma, d, \alpha, L_\nu, \nu, L_\gamma, L_u, T^*, R, v_b, v^*, M)$.

Proof. By a synchronous coupling we have

$$\begin{aligned} d\bar{V}_t^i &= -\lambda (\bar{V}_t^i - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))) dt + \sigma (\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M) dB_t^i, \\ dV_t^i &= -\lambda (V_t^i - \mathcal{P}_{v_b, R}(v_\alpha(\hat{\rho}_t^N))) dt + \sigma (\|\bar{V}_t^i - v_\alpha(\hat{\rho}_t^N)\|_2 \wedge M) dB_t^i, \end{aligned}$$

with coinciding Brownian motions. Moreover, recall that $\text{Law}(\bar{V}_t^i) = \rho_t$ and $\hat{\rho}_t^N = 1/N \sum_{i=1}^N \delta_{V_t^i}$. By Itô's formula we then have

$$\begin{aligned} d\|\bar{V}_t^i - V_t^i\|_2^2 &= \left(-2\lambda \langle \bar{V}_t^i - V_t^i, (\bar{V}_t^i - V_t^i) - (\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - \mathcal{P}_{v_b, R}(v_\alpha(\hat{\rho}_t^N))) \rangle \right. \\ &\quad \left. + \sigma^2 d \left(\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M - \|V_t^i - v_\alpha(\hat{\rho}_t^N)\|_2 \wedge M \right)^2 \right) dt \\ &\quad + 2\sigma \left(\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M - \|V_t^i - v_\alpha(\hat{\rho}_t^N)\|_2 \wedge M \right) (\bar{V}_t^i - V_t^i)^\top dB_t^i, \end{aligned} \quad (21)$$

and after taking the expectation on both sides

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t^i - V_t^i\|_2^2 \right] &= -2\lambda \mathbb{E} \left[\langle \bar{V}_t^i - V_t^i, (\bar{V}_t^i - V_t^i) - (\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - \mathcal{P}_{v_b, R}(v_\alpha(\hat{\rho}_t^N))) \rangle \right] \\ &\quad + \sigma^2 d\mathbb{E} \left[\left(\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M - \|V_t^i - v_\alpha(\hat{\rho}_t^N)\|_2 \wedge M \right)^2 \right] \\ &\leq -2\lambda \mathbb{E} \left[\|\bar{V}_t^i - V_t^i\|_2^2 \right] + \sigma^2 d\mathbb{E} \left[\left\| (\bar{V}_t^i - V_t^i) - (v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)) \right\|_2^2 \right] \\ &\quad + 2\lambda \mathbb{E} \left[\|\bar{V}_t^i - V_t^i\|_2 \left\| \mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - \mathcal{P}_{v_b, R}(v_\alpha(\hat{\rho}_t^N)) \right\|_2 \right] \\ &\leq -2\lambda \mathbb{E} \left[\|\bar{V}_t^i - V_t^i\|_2^2 \right] + 2\lambda \mathbb{E} \left[\|\bar{V}_t^i - V_t^i\|_2 \|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2 \right] \\ &\quad + \sigma^2 d\mathbb{E} \left[\left\| (\bar{V}_t^i - V_t^i) - (v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)) \right\|_2^2 \right]. \end{aligned} \quad (22)$$

Here, let us remark that the last (stochastic) term in (21) disappears after taking the expectation. This is due to $\mathbb{E} \left[\|\bar{V}_t^i - V_t^i\|_2^2 \right] < \infty$, which can be derived from Lemma 8 after noticing that

Lemma 8 also holds for processes V_t^i . Since by Young's inequality it holds

$$2\lambda\mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2 \|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2\right] \leq \lambda\left(\frac{\mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right]}{2} + 2\mathbb{E}\left[\|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right]\right),$$

and

$$\mathbb{E}\left[\left\|\left(\bar{V}_t^i - V_t^i\right) - \left(v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\right)\right\|_2^2\right] \leq 2\mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2 + \|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right],$$

we obtain

$$\begin{aligned} \frac{d}{dt}\mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right] &\leq \left(-\frac{3\lambda}{2} + 2\sigma^2 d\right)\mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right] \\ &\quad + 2(\lambda + \sigma^2 d)\mathbb{E}\left[\|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right] \end{aligned} \quad (23)$$

after inserting the former two inequalities into Equation (22). For the term $\mathbb{E}\left[\|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right]$ we can decompose

$$\mathbb{E}\left[\|v_\alpha(\rho_t) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right] \leq 2\mathbb{E}\left[\|v_\alpha(\rho_t) - v_\alpha(\bar{\rho}_t^N)\|_2^2\right] + 2\mathbb{E}\left[\|v_\alpha(\bar{\rho}_t^N) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right], \quad (24)$$

where we denote

$$\bar{\rho}_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{\bar{V}_t^i}.$$

For the first term in Equation (24), by Lemma 11, we have

$$\mathbb{E}\left[\|v_\alpha(\rho_t) - v_\alpha(\bar{\rho}_t^N)\|_2^2\right] \leq C_0 \frac{1}{N}$$

for some constant C_0 depending on $\lambda, \sigma, d, \alpha, L_\gamma, L_u, T^*, R, v_b, v^*$ and M . For the second term in Equation (24), by combining [11, Lemma 3.2] and Lemma 8, we obtain

$$\mathbb{E}\left[\|v_\alpha(\bar{\rho}_t^N) - v_\alpha(\hat{\rho}_t^N)\|_2^2\right] \leq C_1 \frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right],$$

for some constant C_1 depending on $\lambda, \sigma, d, \alpha, L_u, R$ and M . Combining these estimates we conclude

$$\begin{aligned} \frac{d}{dt} \frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right] &\leq \left(-\frac{3\lambda}{2} + 2\sigma^2 d + 4C_1(\lambda + \sigma^2 d)\right) \frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right] \\ &\quad + 4(\lambda + \sigma^2 d) C_0 \frac{1}{N}. \end{aligned}$$

After an application of Grönwall's inequality and noting that $\bar{V}_0^i = V_0^i$ for all $i = 1, \dots, N$, we have

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\|\bar{V}_t^i - V_t^i\|_2^2\right] \leq 4(\lambda + \sigma^2 d) \frac{C_0}{N} t e^{(-\frac{3\lambda}{2} + 2\sigma^2 d + 4C_1(\lambda + \sigma^2 d))t}. \quad (25)$$

for any $t \in [0, T^*]$. Finally, by Jensen's inequality and letting $t = T^*$, we have

$$\mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^N (V_{T^*}^i - \bar{V}_{T^*}^i)\right\|_2^2\right] \leq \frac{C_{\text{MFA}}}{N}, \quad (26)$$

where the constant C_{MFA} depends on $\lambda, \sigma, d, \alpha, L_u, L_\gamma, T^*, R, v_b, v^*$ and M . \square

In the next lemma we show that the distribution of \bar{V}_t is sub-Gaussian.

Lemma 8. *Let R and M be finite with $R \geq \|v_b - v^*\|_2$. For any $\kappa > 0$, let N satisfy $N \geq (4\sigma^2 M^2)/(\lambda\kappa^2)$. Then, provided that $\mathbb{E} \left[\exp\left(\sum_{i=1}^N \|\bar{V}_0^i - v^*\|_2^2 / (N\kappa^2)\right) \right] < \infty$, it holds*

$$C_\kappa := \sup_{t \in [0, T^*]} \mathbb{E} \left[\exp \left(\frac{\sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2}{N\kappa^2} \right) \right] < \infty, \quad (27)$$

where C_κ depends on $\kappa, \lambda, \sigma, d, R, M$ and T^* , and where

$$d\bar{V}_t^i = -\lambda (\bar{V}_t^i - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))) dt + \sigma (\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M) dB_t^i$$

for $i = 1, \dots, N$ with B_t^i being independent to each other and $\text{Law}(\bar{V}_t^i) = \rho_t$.

Proof. To apply Itô's formula, we need to truncate the function $\exp(\|v\|_2^2/\kappa^2)$ from above. For this, define for $W > 0$ the function

$$G_W(x) := \begin{cases} x & x \in [0, W-1] \\ \frac{1}{16}(x+1-W)^4 - \frac{1}{4}(x+1-W)^3 + x & x \in [W-1, W+1] \\ W & x \in [W+1, \infty) \end{cases}.$$

It is easy to verify that G_W is a \mathcal{C}^2 approximation of the function $x \wedge W$ satisfying $G_W \in \mathcal{C}^2(\mathbb{R}^+)$, $G_W(x) \leq x \wedge W$, $G'_W \in [0, 1]$ and $G''_W \leq 0$.

Since $G_{W, N, \kappa}(t) := \exp(G_W(\sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 / N) / \kappa^2)$ is upper bounded, we can apply Itô's formula to it. We abbreviate $G'_W := G'_W(\sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 / N)$ and $G''_W := G''_W(\sum_{i=1}^N \|\bar{V}_t^i\|_2^2 / N)$ in what follows. With the notation $Y_t := ((\bar{V}_t^1)^\top, \dots, (\bar{V}_t^N)^\top)^\top$, the Nd dimensional process Y_t satisfies $dY_t = -\lambda(Y_t - \overline{\mathcal{P}_{v_b, R}(\rho_t)}) dt + \mathcal{M} dB_t$, where $\overline{\mathcal{P}_{v_b, R}(\rho_t)} = (\mathcal{P}_{v_b, R}(\rho_t)^\top, \dots, \mathcal{P}_{v_b, R}(\rho_t)^\top)^\top$, $\mathcal{M} = \text{diag}(\mathcal{M}_1, \dots, \mathcal{M}_N)$ with $\mathcal{M}_i = \sigma \|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M I_d$ and B_t the Nd dimensional Brownian motion. We then have $G_{W, N, \kappa}(t) = \exp(G_W(\|Y_t\|_2^2 / N) / \kappa^2)$ and

$$\begin{aligned} dG_{W, N, \kappa}(t) &= \sum_{i=1}^N \nabla_{Y_t} G_{W, N, \kappa}(t) dY_t + \frac{1}{2} \text{tr}(\mathcal{M} \nabla_{Y_t, Y_t}^2 G_{W, N, \kappa}(t) \mathcal{M}) dt \\ &= G_{W, N, \kappa}(t) \frac{G'_W}{\kappa^2} \sum_{i=1}^N \left(2 \frac{\bar{V}_t^i - v^*}{N} \right)^\top d\bar{V}_t^i \\ &\quad + \frac{1}{2} G_{W, N, \kappa}(t) \sum_{i=1}^N \left(G'_W \frac{2d}{N\kappa^2} + G''_W \frac{4 \|\bar{V}_t^i - v^*\|_2^2}{N^2 \kappa^2} \right. \\ &\quad \left. + (G'_W)^2 \frac{4 \|\bar{V}_t^i - v^*\|_2^2}{N^2 \kappa^4} \right) (\sigma \|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M)^2 dt. \end{aligned} \quad (28)$$

The first term on the right-hand side of (28) can be expanded as follows

$$\begin{aligned}
G_{W,N,\kappa}(t) \frac{G'_W}{\kappa^2} \sum_{i=1}^N \left(2 \frac{\bar{V}_t^i - v^*}{N} \right)^\top d\bar{V}_t^i &= G_{W,N,\kappa}(t) G'_W \sum_{i=1}^N \left(2 \frac{\bar{V}_t^i - v^*}{N\kappa^2} \right)^\top d\bar{V}_t^i \\
&= G_{W,N,\kappa}(t) G'_W \sum_{i=1}^N \left(2 \frac{\bar{V}_t^i - v^*}{N\kappa^2} \right)^\top \left(-\lambda (\bar{V}_t^i - v^* + v^* - \mathcal{P}_{v_b,R}(\rho_t)) \right) dt + \sigma \left(\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M \right) dB_t^i \\
&= G_{W,N,\kappa}(t) G'_W \left\{ \frac{-2\lambda}{N\kappa^2} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 dt - \frac{2\lambda}{N\kappa^2} \sum_{i=1}^N \langle \bar{V}_t^i - v^*, v^* - \mathcal{P}_{v_b,R}(v_\alpha(\rho_t)) \rangle dt \right. \\
&\quad \left. + 2\sigma \sum_{i=1}^N \left(\|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M \right) \left(\frac{\bar{V}_t^i - v^*}{N\kappa^2} \right)^\top dB_t^i \right\}. \tag{29}
\end{aligned}$$

Notice additionally that

$$\langle \bar{V}_t^i - v^*, v^* - \mathcal{P}_{v_b,R}(v_\alpha(\rho_t)) \rangle \leq \|\bar{V}_t^i - v^*\|_2 \|v^* - \mathcal{P}_{v_b,R}(v_\alpha(\rho_t))\|_2 \leq 2R \|\bar{V}_t^i - v^*\|_2 \tag{30}$$

as v^* and $\mathcal{P}_{v_b,R}(v_\alpha(\rho_t))$ belong to the same ball $B_R(v_b)$ around v_b of radius R . Similarly, we can expand the coefficient of the second term. According to the properties $G'_W \in [0, 1]$ and $G''_W \leq 0$ we can bound it from above yielding

$$\begin{aligned}
\frac{1}{2} G_{W,N,\kappa}(t) \sum_{i=1}^N \left(G'_W \frac{2d}{N\kappa^2} + G''_W \frac{4\|\bar{V}_t^i - v^*\|_2^2}{N^2\kappa^2} + (G'_W)^2 \frac{4\|\bar{V}_t^i - v^*\|_2^2}{N^2\kappa^4} \right) (\sigma \|\bar{V}_t^i - v_\alpha(\rho_t)\|_2 \wedge M)^2 \\
\leq G_{W,N,\kappa}(t) G'_W \frac{\sigma^2 M^2 d}{\kappa^2} + G_{W,N,\kappa}(t) (G'_W)^2 \frac{2\sigma^2 M^2}{N^2\kappa^4} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 \\
\leq G_{W,N,\kappa}(t) G'_W \frac{\sigma^2 M^2 d}{\kappa^2} + G_{W,N,\kappa}(t) G'_W \frac{2\sigma^2 M^2}{N^2\kappa^4} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2. \tag{31}
\end{aligned}$$

By taking expectations in (28) and combining it with (29), (30) and (31), we obtain

$$\begin{aligned}
\frac{d}{dt} \mathbb{E} [G_{W,N,\kappa}(t)] &\leq \mathbb{E} \left[G_{W,N,\kappa}(t) G'_W \left(\frac{-2\lambda}{N\kappa^2} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 + \frac{4R\lambda}{N\kappa^2} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2 \right) \right. \\
&\quad \left. + G_{W,N,\kappa}(t) G'_W \frac{\sigma^2 M^2 d}{\kappa^2} + G_{W,N,\kappa}(t) G'_W \frac{2\sigma^2 M^2}{N^2\kappa^4} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 \right]
\end{aligned}$$

Rearranging the former yields

$$\begin{aligned}
\frac{d}{dt} \mathbb{E} [G_{W,N,\kappa}(t)] &\leq \mathbb{E} \left[G_{W,N,\kappa}(t) G'_W \left(\left(\frac{4\lambda R}{N\kappa^2} \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2 \right) + \frac{\sigma^2 M^2 d}{\kappa^2} \right) \right. \\
&\quad \left. - \left(\frac{2\lambda}{N\kappa^2} - \frac{2\sigma^2 M^2}{N^2\kappa^4} \right) \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 \right], \tag{32}
\end{aligned}$$

Since by Young's inequality, it holds $4R \|\bar{V}_t^i - v^*\|_2 \leq 4R^2 + \|\bar{V}_t^i - v^*\|_2^2$, we can continue

Estimate (32) by

$$\begin{aligned} \frac{d}{dt} \mathbb{E} [G_{W,N,\kappa}(t)] &\leq \mathbb{E} \left[G_{W,N,\kappa}(t) G'_W \left(\frac{\sigma^2 M^2 d + 4\lambda R^2}{\kappa^2} - \left(\frac{\lambda}{N\kappa^2} - \frac{2\sigma^2 M^2}{N^2 \kappa^4} \right) \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 \right) \right] \\ &\leq \mathbb{E} \left[G_{W,N,\kappa}(t) G'_W \left(-A \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 + B \right) \right] \end{aligned} \quad (33)$$

with $A := \frac{\lambda}{N\kappa^2} - \frac{2\sigma^2 M^2}{N^2 \kappa^4}$ and $B := \frac{\sigma^2 M^2 d + 4\lambda R^2}{\kappa^2}$. Now, if $\sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 \geq (B-1)/A$, we have

$$G_{W,N,\kappa}(t) G'_W \left(-A \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 + B \right) \leq 0,$$

while, if $\sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 \leq (B-1)/A$, we have

$$G_{W,N,\kappa}(t) G'_W \left(-A \sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2 + B \right) \leq B e^{\frac{B-1}{N\kappa^2 A}}.$$

Thus the latter inequality always holds true and consequently we have with (33)

$$\frac{d}{dt} \mathbb{E} [G_{W,N,\kappa}(t)] \leq B e^{\frac{B-1}{N\kappa^2 A}},$$

which gives after integration

$$\begin{aligned} \mathbb{E} [G_{W,N,\kappa}(t)] &\leq \mathbb{E} [G_{W,N,\kappa}(0)] + B e^{\frac{B-1}{N\kappa^2 A}} t \\ &\leq \mathbb{E} \left[\exp \left(\frac{\sum_{i=1}^N \|\bar{V}_0^i - v^*\|_2^2}{N\kappa^2} \right) \right] + B e^{\frac{B-1}{N\kappa^2 A}} t. \end{aligned}$$

Letting $W \rightarrow \infty$, we eventually obtain

$$\mathbb{E} \left[\exp \left(\frac{\sum_{i=1}^N \|\bar{V}_t^i - v^*\|_2^2}{N\kappa^2} \right) \right] \leq \mathbb{E} \left[\exp \left(\frac{\sum_{i=1}^N \|\bar{V}_0^i - v^*\|_2^2}{N\kappa^2} \right) \right] + B e^{\frac{B-1}{N\kappa^2 A}} t < \infty, \quad (34)$$

provided that $\mathbb{E} \left[\exp \left(\frac{\sum_{i=1}^N \|\bar{V}_0^i - v^*\|_2^2}{N\kappa^2} \right) \right] < \infty$.

If $N \geq (4\sigma^2 M^2)/(\lambda\kappa^2)$, we have

$$\frac{B-1}{N\kappa^2 A} \leq \frac{B}{N\kappa^2 A} = \frac{N(\sigma^2 M^2 d + 4\lambda R^2)}{\lambda N\kappa^2 - 2\sigma^2 M^2} \leq C(\kappa, \lambda, \sigma, M, R, d).$$

Thus, C_κ is upper bounded and independent of N . \square

Remark 9. The sub-Gaussianity of \bar{V}_t follows from Lemma 8 by noticing that the statement can be applied in the setting $N = 1$ when choosing κ sufficiently large.

Remark 10. In Lemma 8, as the number of particles N increases, the condition for κ to ensure $C_\kappa < \infty$ becomes more relaxed. Specifically, the value of κ can be as small as one needs as N increases. This phenomenon can be easily understood by considering the limit as N approaches infinity. In this case, C_κ tends to $\sup_{t \in [0, T^*]} \exp \left(\mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] / \kappa^2 \right)$. Therefore, as one shows an upper bound on the second moment of \bar{V}_t , it becomes evident that C_κ remains finite as N tends to infinity.

With the help of Lemma 8, we can now prove the following lemma.

Lemma 11. *Let $f \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 and A3. Then, for any $t \in [0, T^*]$, M and R with $R \geq \|v_b - v^*\|_2$ finite, and N satisfying $N \geq (16\alpha L_\gamma \sigma^2 M^2)/\lambda$, we have*

$$\mathbb{E} \left[\|v_\alpha(\rho_t) - v_\alpha(\bar{\rho}_t^N)\|_2^2 \right] \leq \frac{C_0}{N}, \quad (35)$$

where $C_0 := C_0(\lambda, \sigma, d, \alpha, L_\gamma, L_u, T^*, R, v_b, v^*, M)$.

Proof. Without loss of generality, we assume $v^* = 0$ and recall that we assumed $\underline{f} = 0$ in the proofs as of Remark 6. We have

$$\begin{aligned} \mathbb{E} \left[\|v_\alpha(\rho_t) - v_\alpha(\bar{\rho}_t^N)\|_2^2 \right] &= \mathbb{E} \left[\left\| \frac{\frac{1}{N} \sum_{i=1}^N \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)}}{\frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)}} - \frac{\int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v)}{\int_{\mathbb{R}^d} e^{-\alpha f(v)} d\rho_t(v)} \right\|_2^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| \frac{1}{\frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)}} \left(\frac{1}{N} \sum_{i=1}^N \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v) \right) \right\|_2^2 \right] \\ &\quad + 2\mathbb{E} \left[\left\| \frac{v_\alpha(\rho_t)}{\frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)}} \left(\frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} e^{-\alpha f(v)} d\rho_t(v) \right) \right\|_2^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| e^{\alpha \frac{1}{N} \sum_{i=1}^N f(\bar{V}_t^i)} \left(\frac{1}{N} \sum_{i=1}^N \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v) \right) \right\|_2^2 \right] \\ &\quad + 2\|v_\alpha(\rho_t)\|_2^2 \mathbb{E} \left[\left\| e^{\alpha \frac{1}{N} \sum_{i=1}^N f(\bar{V}_t^i)} \left(\frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} e^{-\alpha f(v)} d\rho_t(v) \right) \right\|_2^2 \right] \\ &\leq 2T_1 T_2 + 2\|v_\alpha(\rho_t)\|_2^2 T_1 T_3, \end{aligned} \quad (36)$$

where we defined

$$\begin{aligned} T_1 &:= \left(\mathbb{E} \left[e^{4\alpha \frac{1}{N} \sum_{i=1}^N f(\bar{V}_t^i)} \right] \right)^{\frac{1}{2}}, \\ T_2 &:= \left(\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v) \right\|_2^4 \right] \right)^{\frac{1}{2}}, \\ T_3 &:= \left(\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} e^{-\alpha f(v)} d\rho_t(v) \right\|_2^4 \right] \right)^{\frac{1}{2}}. \end{aligned}$$

In the following, we upper bound the terms T_1, T_2 and T_3 separately. Firstly, recall that by Lemma 8 we have for $t \in [0, T^*]$ that

$$\mathbb{E} \left[\exp \left(\frac{\sum_{i=1}^N \|\bar{V}_t^i\|_2^2}{N\kappa^2} \right) \right] \leq C_\kappa < \infty, \quad (37)$$

where C_κ only depends on $\kappa, \lambda, \sigma, d, R, M$ and T^* . With this,

$$\begin{aligned}
T_1^2 &= \mathbb{E} \left[\exp \left(4\alpha \frac{1}{N} \sum_{i=1}^N f(\bar{V}_t^i) \right) \right] \leq \mathbb{E} \left[\exp \left(4\alpha \frac{1}{N} \sum_{i=1}^N L_\gamma (1 + \|\bar{V}_t^i\|_2^{1+\gamma}) \right) \right] \\
&\leq e^{4\alpha L_\gamma} \mathbb{E} \left[\exp \left(4\alpha L_\gamma \frac{1}{N} \sum_{i=1}^N \|\bar{V}_t^i\|_2^{1+\gamma} \right) \right] \\
&\leq e^{8\alpha L_\gamma} \mathbb{E} \left[\exp \left(4\alpha L_\gamma \frac{1}{N} \sum_{i=1}^N \|\bar{V}_t^i\|_2^2 \right) \right] \\
&= e^{8\alpha L_\gamma} \mathbb{E} \left[\exp \left(\frac{1}{\kappa^2} \frac{1}{N} \sum_{i=1}^N \|\bar{V}_t^i\|_2^2 \right) \right] \\
&\leq e^{8\alpha L_\gamma} C_\kappa \Big|_{\kappa = \frac{1}{2\sqrt{\alpha L_\gamma}}},
\end{aligned}$$

where we set $\kappa^2 = 1/(4\alpha L_\gamma)$ in the next-to-last step and where N should satisfy $N \geq (16\alpha L_\gamma \sigma^2 M^2)/\lambda$. Secondly, we have

$$\begin{aligned}
\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v) \right\|_2^4 \right] &= \frac{1}{N^4} \mathbb{E} \left[\sum_{i_1, i_2, i_3, i_4 \in \{1, \dots, N\}} \langle \bar{Z}_t^{i_1}, \bar{Z}_t^{i_2} \rangle \langle \bar{Z}_t^{i_3}, \bar{Z}_t^{i_4} \rangle \right] \\
&\leq \frac{4! L_u^4}{N^2},
\end{aligned}$$

where $(\bar{Z}_t^i := \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v))_{i=1, \dots, N}$ are i.i.d. and have zero mean. Thus,

$$T_2 = \left(\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_t^i e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} v e^{-\alpha f(v)} d\rho_t(v) \right\|_2^4 \right] \right)^{\frac{1}{2}} \leq \frac{5L_u^2}{N}.$$

Similarly, we can derive

$$T_3 = \left(\mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N e^{-\alpha f(\bar{V}_t^i)} - \int_{\mathbb{R}^d} e^{-\alpha f(v)} d\rho_t(v) \right\|_2^4 \right] \right)^{\frac{1}{2}} \leq \frac{5}{N}.$$

Collecting the bounds for the terms T_1, T_2 and T_3 and inserting them in (36), we obtain

$$\mathbb{E} \left[\|v_\alpha(\rho_t) - v_\alpha(\bar{\rho}_t)\|_2^2 \right] \leq 10e^{6\alpha L_\gamma} C_\kappa^{\frac{1}{2}} \Big|_{\kappa = \frac{1}{2\sqrt{\alpha L_\gamma}}} \left(L_u^2 + \sup_{t \in [0, T^*]} \|v_\alpha(\rho_t)\|_2^2 \right) \frac{1}{N}. \quad (38)$$

Since by Lemmas 14, 16 and 17, we know that $\|v_\alpha(\rho_t)\|_2$ can be uniformly bounded by a constant depending on $\alpha, \lambda, \sigma, d, R, v_b, v^*, M, L_\nu$ and ν (see in particular Equation (48) that combines the aforementioned lemmas), we can conclude (38) with

$$\mathbb{E} \left[\|v_\alpha(\rho_t) - v_\alpha(\bar{\rho}_t)\|_2^2 \right] \leq \frac{C_0}{N} \quad (39)$$

for some constant C_0 depends on $\lambda, \sigma, d, \alpha, L_\nu, \nu, L_\gamma, L_u, T^*, R, v_b, v^*$ and M . \square

3.2.2 Upper Bound for the Third Term in (19)

In this section, we bound Term III of the error decomposition (19). Before stating the main result of this section, Proposition 15, we first need to provide two auxiliary lemmas, Lemma 12 and Lemma 14.

Lemma 12. *Let $R, M \in (0, \infty)$. Then it holds*

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &\leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \left(\|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 + \|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right) \\ &\quad + \sigma^2 M^2 d. \end{aligned} \quad (40)$$

If further $\lambda \geq 2\sigma^2 d$, we have

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &\leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \left(\|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 + \|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right). \end{aligned} \quad (41)$$

Proof. By Itô's formula, we have

$$\begin{aligned} d \|\bar{V}_t - v^*\|_2^2 &= 2 (\bar{V}_t - v^*)^\top d\bar{V}_t + \sigma^2 d \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \wedge M^2 \right) dt \\ &= -2\lambda \langle \bar{V}_t - v^*, \bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) \rangle dt + 2\sigma \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2 \wedge M \right) (\bar{V}_t - v^*)^\top dB_t \\ &\quad + \sigma^2 d \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \wedge M^2 \right) dt \\ &= -\lambda \left[\|\bar{V}_t - v^*\|_2^2 + \|\bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 - \|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 \right] dt \\ &\quad + 2\sigma \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2 \wedge M \right) (\bar{V}_t - v^*)^\top dB_t + \sigma^2 d \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \wedge M^2 \right) dt, \end{aligned}$$

which, after taking the expectation on both sides, yields

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &= -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 - \lambda \mathbb{E} \left[\|\bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right] \\ &\quad + \sigma^2 d \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \wedge M^2 \right]. \end{aligned} \quad (42)$$

For the term $\mathbb{E} \left[\|\bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right]$, we notice that

$$\begin{aligned} \mathbb{E} \left[\|\bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right] &= \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right] + \mathbb{E} \left[\|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right] \\ &\quad + 2\mathbb{E} \left[\langle \bar{V}_t - v_\alpha(\rho_t), v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) \rangle \right] \\ &\geq \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right] + \mathbb{E} \left[\|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right] \\ &\quad - \left(\frac{1}{2} \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right] + 2\mathbb{E} \left[\|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right] \right) \\ &= \frac{1}{2} \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right] - \mathbb{E} \left[\|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right], \end{aligned}$$

which, inserted into Equation (42), allows to derive

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &\leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \left(\|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 + \|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right) \\ &\quad - \frac{1}{2} \lambda \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right] + \sigma^2 d \left(\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \wedge M^2 \right). \end{aligned}$$

From this we get for any λ and σ that

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &\leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \left(\|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 + \|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right) \\ &\quad + \sigma^2 M^2 d. \end{aligned} \quad (43)$$

as well as

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &\leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \left(\|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 + \|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right) \\ &\quad + \left(-\frac{1}{2}\lambda + \sigma^2 d \right) \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right]. \end{aligned} \quad (44)$$

If $\lambda \geq 2\sigma^2 d$, by Equation (44), we get

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &\leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \left(\|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 + \|v_\alpha(\rho_t) - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2 \right). \end{aligned} \quad (45)$$

□

Remark 13. When $R = M = \infty$, we can show

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] &= -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \\ &\quad + \lambda \|v_\alpha(\rho_t) - v^*\|_2^2 - (\lambda - \sigma^2 d) \mathbb{E} \left[\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \right]. \end{aligned}$$

If further $\lambda \geq \sigma^2 d$, we have

$$\frac{d}{dt} \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] \leq -\lambda \mathbb{E} \left[\|\bar{V}_t - v^*\|_2^2 \right] + \lambda \|v_\alpha(\rho_t) - v^*\|_2^2.$$

This differs from [23, Lemma 18].

The next result is a quantitative version of the Laplace principle as established in [23, Proposition 21].

Lemma 14. For any $r > 0$, define $f_r := \sup_{v \in B_r(v^*)} f(v)$. Then, under the inverse continuity condition A2, for any $r \in (0, R_0]$ and $q > 0$ such that $q + f_r \leq f_\infty$, it holds

$$\|v_\alpha(\rho) - v^*\|_2 \leq \frac{(q + f_r)^\nu}{L_\nu} + \frac{\exp(-\alpha q)}{\rho(B_r(v^*))} \int \|v - v^*\|_2 d\rho(v) \quad (46)$$

With the above preparation, we can now upper bound Term III. We have by Jensen's inequality

$$III = \mathbb{E} \left[\left\| \frac{1}{N} \sum_{i=1}^N \bar{V}_{T^*}^i - v^* \right\|_2^2 \right] \leq \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[\|\bar{V}_{T^*}^i - v^*\|_2^2 \right], \quad (47)$$

i.e., it is enough to upper bound $\mathbb{E} \left[\|\bar{V}_{T^*} - v^*\|_2^2 \right]$, which is the content of the next statement.

Proposition 15. Let $f \in \mathcal{C}(\mathbb{R}^d)$ satisfy [A1](#), [A2](#) and [A3](#). Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ with $v^* \in \text{supp}(\rho_0)$. Fix any $\epsilon \in (0, W_2^2(\rho_0, \delta_{v^*}))$ and define the time horizon

$$T^* := \frac{1}{\lambda} \log \left(\frac{2W_2^2(\rho_0, \delta_{v^*})}{\epsilon} \right).$$

Moreover, let $R \in (\|v_b - v^*\|_2 + \sqrt{\epsilon/2}, \infty)$, $M \in (0, \infty)$ and $\lambda, \sigma > 0$ be such that $\lambda \geq 2\sigma^2 d$ or $\sigma^2 M^2 d = \mathcal{O}(\epsilon)$. Then we can choose α sufficiently large, depending on $\lambda, \sigma, d, T^*, R, v_b, M, \epsilon$ and properties of f , such that $\mathbb{E} [\|\bar{V}_{T^*} - v^*\|_2^2] = \mathcal{O}(\epsilon)$.

Proof. We only prove the case $\lambda \geq 2\sigma^2 d$ in detail. The case $\sigma^2 M^2 d = \mathcal{O}(\epsilon)$ follows similarly.

According to [Lemmas 14](#) and [17](#), we have

$$\begin{aligned} \|v_\alpha(\rho_t) - v^*\|_2 &\leq \frac{(q + f_r)^\nu}{L_\nu} + \frac{\exp(-\alpha q)}{\rho_t(B_r(v^*))} \mathbb{E} [\|\bar{V}_t - v^*\|_2] \\ &\leq \frac{(q + f_r)^\nu}{L_\nu} + \exp(-\alpha q) C_2 C_3, \end{aligned} \tag{48}$$

where $C_2 := (\exp\{q'T^*\})/C_4 < \infty$, q' and C_4 are from [Lemma 17](#), and where, as of [Lemma 16](#), $C_3 := \sup_{[0, T^*]} \mathbb{E} [\|\bar{V}_t - v^*\|_2] < \infty$. In what follows, let us deal with the two terms on the right-hand side of (48). For the term $(q + f_r)^\nu/L_\nu$, let $q = f_r$. Then by [A2](#) and [A3](#), we can choose proper r , such that $2(L_\nu r)^{1/\nu} \leq 2f_r \leq f_\infty$. Further by [A3](#), we have

$$\frac{(q + f_r)^\nu}{L_\nu} = \frac{(2f_r)^\nu}{L_\nu} \leq \frac{(2L_\gamma)^\nu r^{(1+\gamma)\nu}}{L_\nu},$$

so if

$$r < r_0 := \min \left\{ \left(\frac{\epsilon}{8} \right)^{\frac{1}{2(1+\gamma)\nu}} \left(\frac{L_\nu}{(2L_\gamma)^\nu} \right)^{\frac{1}{(1+\gamma)\nu}}, \sqrt{\frac{\epsilon}{2}} \right\},$$

we can bound

$$\frac{(q + f_r)^\nu}{L_\nu} = \frac{(2f_r)^\nu}{L_\nu} \leq \frac{\sqrt{\epsilon}}{2\sqrt{2}}.$$

For term $\exp(-\alpha q) C_2 C_3$, we can choose α large enough such that

$$\exp(-\alpha q) C_2 C_3 \leq \frac{\sqrt{\epsilon}}{2\sqrt{2}}.$$

With these choices of r and α and by integrating them into [Equation \(48\)](#), we obtain

$$\|v_\alpha(\rho_t) - v^*\|_2^2 < \frac{\epsilon}{2},$$

for all $t \in [0, T^*]$, and thus

$$\|v_\alpha(\rho_t) - v_b\|_2 \leq \|v_\alpha(\rho_t) - v^*\|_2 + \|v^* - v_b\|_2 \leq \sqrt{\frac{\epsilon}{2}} + \|v^* - v_b\|_2 \leq R.$$

Consequently, by [Lemma 12](#), we have

$$\begin{aligned} \frac{d}{dt} \mathbb{E} [\|\bar{V}_t - v^*\|_2^2] &\leq -\lambda \left(\mathbb{E} [\|\bar{V}_t - v^*\|_2^2] - \|v_\alpha(\rho_t) - v^*\|_2^2 \right) \\ &\leq -\lambda \left(\mathbb{E} [\|\bar{V}_t - v^*\|_2^2] - \frac{\epsilon}{2} \right), \end{aligned}$$

since now $\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) = v_\alpha(\rho_t)$. Finally by Grönwall's inequality, $\mathbb{E} [\|\bar{V}_{T^*} - v^*\|_2^2] \leq \epsilon$. \square

Lemma 16. *Let $\|v_b - v^*\|_2 < R < \infty$ and $0 < M < \infty$. Then it holds*

$$\sup_{t \in [0, T^*]} \mathbb{E} [\|\bar{V}_t - v^*\|_2] \leq \sqrt{\max \left\{ \mathbb{E} [\|\bar{V}_0 - v^*\|_2^2], \lambda R^2 + \sigma^2 M^2 d \right\}}. \quad (49)$$

Proof. By Equation (42) we have

$$\begin{aligned} \frac{d}{dt} \mathbb{E} [\|\bar{V}_t - v^*\|_2^2] &\leq -\lambda \mathbb{E} [\|\bar{V}_t - v^*\|_2^2] \\ &\quad + \lambda \|\mathcal{P}_{v_b, R}(v_\alpha(\rho_t)) - v^*\|_2^2 - \lambda \mathbb{E} [\|\bar{V}_t - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))\|_2^2] \\ &\quad + \sigma^2 d \mathbb{E} [\|\bar{V}_t - v_\alpha(\rho_t)\|_2^2 \wedge M^2] \\ &\leq -\lambda \mathbb{E} [\|\bar{V}_t - v^*\|_2^2] + \lambda R^2 + \sigma^2 M^2 d, \end{aligned}$$

yielding

$$\mathbb{E} [\|\bar{V}_t - v^*\|_2^2] \leq \max \left\{ \mathbb{E} [\|\bar{V}_0 - v^*\|_2^2], \lambda R^2 + \sigma^2 M^2 d \right\},$$

after an application of Grönwall's inequality for any $t \geq 0$. \square

Lemma 17. *For any $M \in (0, \infty)$, $\tau \geq 1$, $r > 0$ and $R \in (\|v_b - v^*\|_2 + r, \infty)$ it holds*

$$\rho_t(B_r(v^*)) \geq C_4 \exp(-q't) > 0,$$

where

$$C_4 := \int_{B_r(v^*)} 1 + (\tau - 1) \left\| \frac{v - v^*}{r} \right\|_2^\tau - \tau \left\| \frac{v - v^*}{r} \right\|_2^{\tau-1} d\rho_0(v)$$

and where q' depends on $\tau, \lambda, \sigma, d, r, R, v_b$ and M .

Proof. Recall that the law ρ_t of \bar{V}_t satisfies the Fokker-Planck equation

$$\partial_t \rho_t = \lambda \operatorname{div}((v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))) \rho_t) + \frac{\sigma^2}{2} \Delta \left((\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2) \rho_t \right).$$

Let us first define for $\tau \geq 1$ the test function

$$\phi_r^\tau(v) := \begin{cases} 1 + (\tau - 1) \left\| \frac{v}{r} \right\|_2^\tau - \tau \left\| \frac{v}{r} \right\|_2^{\tau-1}, & \|v\|_2 \leq r, \\ 0, & \text{else,} \end{cases} \quad (50)$$

for which it is easy to verify that $\phi_r^\tau \in C_c^1(\mathbb{R}^d, [0, 1])$. Since $\operatorname{Im} \phi_r^\tau \subset [0, 1]$, we have $\rho_t(B_r(v^*)) \geq \int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_t(v)$. To lower bound $\rho_t(B_r(v^*))$, it is thus sufficient to establish a lower bound on $\int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_t(v)$. By Green's formula

$$\begin{aligned} \frac{d}{dt} \int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_t(v) &= -\lambda \int_{B_r(v^*)} \langle v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t)), \nabla \phi_r^\tau(v - v^*) \rangle d\rho_t(v) \\ &\quad + \frac{\sigma^2}{2} \int_{B_r(v^*)} (\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2) \Delta \phi_r^\tau(v - v^*) d\rho_t(v) \\ &= \tau(\tau - 1) \int_{B_r(v^*)} \frac{\|v - v^*\|_2^{\tau-3}}{r^{\tau-3}} \left(\left(1 - \frac{\|v - v^*\|_2}{r} \right) \left(\lambda \left\langle \frac{v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))}{r}, \frac{v - v^*}{r} \right\rangle \right) \right. \\ &\quad \left. - \frac{\sigma^2}{2} (d + \tau - 2) \frac{\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2}{r^2} \right) + \frac{\sigma^2}{2} \frac{\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2}{r^2} d\rho_t(v). \end{aligned}$$

For simplicity, let us abbreviate

$$\Theta := \left(1 - \frac{\|v - v^*\|_2}{r}\right) \left(\lambda \left\langle \frac{v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))}{r}, \frac{v - v^*}{r} \right\rangle - \frac{\sigma^2}{2} (d + \tau - 2) \frac{\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2}{r^2}\right) + \frac{\sigma^2}{2} \frac{\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2}{r^2}.$$

We can choose ϵ_1 small enough, depending on τ and d , such that when $\|v - v^*\|_2/r > 1 - \epsilon_1$, we have

$$\begin{aligned} \Theta &= \left(1 - \frac{\|v - v^*\|_2}{r}\right) \lambda \left\langle \frac{v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))}{r}, \frac{v - v^*}{r} \right\rangle \\ &\quad + \left(\frac{\sigma^2}{2} - \left(1 - \frac{\|v - v^*\|_2}{r}\right) \frac{\sigma^2}{2} (d + \tau - 2)\right) \frac{\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2}{r^2} \\ &\geq \left(1 - \frac{\|v - v^*\|_2}{r}\right) \lambda \left\langle \frac{v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t))}{r}, \frac{v - v^*}{r} \right\rangle + \frac{\sigma^2}{3} \frac{\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2}{r^2}, \end{aligned}$$

where the last inequality works if $\|v - v^*\|_2/r \geq 1 - 1/(6(d + \tau - 2))$.

If $v_\alpha(\rho_t) \notin B_R(v_b)$, we have $|\langle v - \mathcal{P}_{v_b, R}(v_\alpha(\rho_t)), v - v^* \rangle|/r^2 \leq C(r, R, v_b)$ and, since $R > \|v_b - v^*\|_2 + r$, $(\|v - v_\alpha(\rho_t)\|_2^2 \wedge M^2)/r^2 \geq C(r, M, R, v_b)$, which allows to choose ϵ_2 small enough, depending on $\lambda, r, \sigma, R, v_b$ and M , such that $\Theta > 0$ when $\|v - v^*\|_2/r > 1 - \min\{\epsilon_1, \epsilon_2\}$.

If $v_\alpha(\rho_t) \in B_R(v_b)$ and $\|v - v_\alpha(\rho_t)\|_2 \leq M$, we have by Lemma 18

$$\begin{aligned} \Theta &\geq \left(1 - \frac{\|v - v^*\|_2}{r}\right) \lambda \left\langle \frac{v - v_\alpha(\rho_t)}{r}, \frac{v - v^*}{r} \right\rangle + \frac{\sigma^2}{3} \frac{\|v - v_\alpha(\rho_t)\|_2^2}{r^2} \\ &= \left(\frac{\sigma^2}{3} + \left(1 - \frac{\|v - v^*\|_2}{r}\right) \lambda\right) \frac{\|v - v^*\|_2^2}{r^2} + \frac{\sigma^2}{3} \frac{\|v_\alpha(\rho_t) - v^*\|_2^2}{r^2} \\ &\quad - \left(\frac{2\sigma^2}{3} + \left(1 - \frac{\|v - v^*\|_2}{r}\right) \lambda\right) \left\langle \frac{v_\alpha(\rho_t) - v^*}{r}, \frac{v - v^*}{r} \right\rangle \\ &\geq 0, \end{aligned}$$

when $\|v - v^*\|_2/r \in [1 - 2\sigma^2/(3\lambda), 1]$.

If $v_\alpha(\rho_t) \in B_R(v_b)$ and $\|v - v_\alpha(\rho_t)\|_2 > M$, we have

$$\Theta \geq \left(1 - \frac{\|v - v^*\|_2}{r}\right) C(\lambda, r, R, v_b) + \frac{\sigma^2}{3} M^2,$$

i.e., we can choose ϵ_3 small enough, depending on $\lambda, r, \sigma, R, v_b$ and M , such that $\Theta \geq 0$ when $\|v - v^*\|_2/r > 1 - \min\{\epsilon_1, \epsilon_2, \epsilon_3, 2\sigma^2/3\lambda\}$.

Combining the cases from above, we conclude that $\Theta \geq 0$ when $\|v - v^*\|_2/r \geq 1 - \min\{\epsilon_1, \epsilon_2, \epsilon_3, 2\sigma^2/3\lambda\}$. On the other hand, when $\|v - v^*\|_2/r \leq 1 - \min\{\epsilon_1, \epsilon_2, \epsilon_3, 2\sigma^2/3\lambda\}$, we have

$$\tau(\tau - 1) \frac{\|v - v^*\|_2^{\tau-3}}{r^{\tau-3}} \Theta = \tau(\tau - 1) \frac{\|v - v^*\|_2^{\tau-3}}{r^{\tau-3}} \frac{\Theta}{\phi_r^\tau(v)} \phi_r^\tau(v - v^*) \geq -C_5 \phi_r^\tau(v - v^*)$$

for some constant C_5 depending on $r, R, M, v_b, \lambda, \sigma, d$ and τ , since $|\Theta|$ is upper bounded and $\phi_r^\tau(v - v^*) \geq \phi_r^\tau((1 - \min\{\epsilon_1, \epsilon_2, \epsilon_3, 2\sigma^2/3\lambda\})r) > 0$ for any v satisfies $\|v - v^*\|_2/r \leq 1 - \min\{\epsilon_1, \epsilon_2, \epsilon_3, 2\sigma^2/3\lambda\}$.

All in all we have

$$\frac{d}{dt} \int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_t(v) \geq -q' \int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_t(v),$$

where $q' := \max\{C_5, 0\}$. By Grönwall's inequality, we thus have

$$\rho_t(B_r(v^*)) \geq \int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_t(v) \geq e^{-q't} \int_{B_r(v^*)} \phi_r^\tau(v - v^*) d\rho_0(v),$$

which concludes the proof. \square

Lemma 18. *Let $a, b > 0$. Then we have*

$$(a + b(1 - x))x^2 + ay^2 - (2a + b(1 - x))xy \geq 0,$$

for any $x \in [1 - 2a/b, 1] \cap (0, \infty)$ and $y \geq 0$.

Proof. For $y = 0$, this is true. For $y > 0$, divide both side by ay^2 and denote $c = b/a$. Then the lemma is equivalent to showing $(1 + c(1 - x))(x/y)^2 - (2 + c(1 - x))x/y + 1 \geq 0$, i.e., it is enough to show $\min_{r \geq 0} (1 + c(1 - x))r^2 - (2 + c(1 - x))r + 1 \geq 0$, when $x \in [1 - 2/c, 1]$. We have

$$\arg \min_r (1 + c(1 - x))r^2 - (2 + c(1 - x))r + 1 = \frac{2 + c(1 - x)}{2 + 2c(1 - x)},$$

and thus

$$\begin{aligned} & \min_{r \geq 0} (1 + c(1 - x))r^2 - (2 + c(1 - x))r + 1 \\ &= (1 + c(1 - x)) \left(\frac{2 + c(1 - x)}{2 + 2c(1 - x)} \right)^2 - (2 + c(1 - x)) \frac{2 + c(1 - x)}{2 + 2c(1 - x)} + 1 \\ &= -\frac{1}{2} \frac{(2 + c(1 - x))^2}{2 + 2c(1 - x)} + 1 \geq 0, \end{aligned}$$

when $x \in [1 - 2/c, 1]$. This finishes the proof. \square

4 Numerical Experiments

In this section we numerically demonstrate the benefit of using CBO with truncated noise. For isotropic [11, 23, 40] and anisotropic noise [13, 24], we compare the CBO method with truncation $M = 1$ to standard CBO for several benchmark problems in optimization, which are summarized in Table 1.

Name	Objective function f	v^*	f
Ackley	$-20 \exp\left(-0.2\sqrt{\frac{1}{d} \sum_{i=1}^d v_i^2}\right) - \exp\left(\frac{1}{d} \sum_{i=1}^d \cos(2\pi v_i)\right) + 20 + e$	$(0, \dots, 0)$	0
Griewank	$1 + \sum_{i=1}^d \frac{v_i^2}{4000} - \prod_{i=1}^d \cos\left(\frac{v_i}{i}\right)$	$(0, \dots, 0)$	0
Rastrigin	$10d + \sum_{i=1}^d [v_i^2 - 10 \cos(2\pi v_i)]$	$(0, \dots, 0)$	0
Alpine	$10 \sum_{i=1}^d \ (v_i - v_i^*) \sin(10(v_i - v_i^*)) - 0.1(v_i - v_i^*)\ _2$	$(0, \dots, 0)$	0
Salomon	$1 - \cos\left(200\pi\sqrt{\sum_{i=1}^d v_i^2}\right) + 10\sqrt{\sum_{i=1}^d v_i^2}$	$(0, \dots, 0)$	0

Table 1: Benchmark test functions

In the subsequent tables we report comparison results for the two methods for the different benchmark functions as well as different numbers of particles N and, potentially, different

numbers of steps K . Throughout, we set $v_b = 0$ and $R = \infty$, which is out of convenience. Any sufficiently large but finite choice for R yields identical results.

The success criterion is defined by achieving the condition $\left\| \frac{1}{N} \sum_{i=1}^N V_{K,\Delta t}^i - v^* \right\|_2 \leq 0.1$, which ensures that the algorithm has reached the basin of attraction of the global minimizer. The success rate is averaged over 1000 runs.

Isotropic Case. Let $d = 15$. In the case of isotropic noise, we always set $\lambda = 1$, $\sigma = 0.3$, $\alpha = 10^5$ and step-size $\Delta t = 0.02$. The initial positions $(V_0^i)_{i=1,\dots,N}$ are sampled i.i.d. from $\rho_0 = \mathcal{N}(0, I_d)$. In Table 2 we report results comparing the isotropic CBO method with truncation $M = 1$ and the original isotropic CBO method [11, 23, 40] ($M = +\infty$) for the Ackley, Griewank and Salomon function. Each algorithm is run for $K = 200$ steps.

Number of steps $K = 200$						
Test function	M	$N = 150$	$N = 300$	$N = 600$	$N = 900$	$N = 1200$
Ackley	1	0.978	0.999	1	1	1
	$+\infty$	0.001	0.056	0.478	0.824	0.935
Griewank	1	0.060	0.188	0.5013	0.671	0.791
	$+\infty$	0	0	0.010	0.013	0.032
Salomon	1	0.970	1	1	1	1
	$+\infty$	0.005	0.068	0.603	0.909	0.979

Table 2: For the 15-dimensional Ackley and Salomon function, the CBO method with truncation ($M = 1$) is able to locate the global minimum using only $N = 300$ particles. In comparison, even with a larger number of particles (up to $N = 1200$), the original CBO method ($M = +\infty$) cannot achieve a flawless success rate. In the case of the Griewank function, the original CBO method ($M = +\infty$) exhibits a quite low success rate, even when utilizing $N = 1200$ particles. Contrarily, in the same setting, the CBO method with truncation ($M = 1$) achieves a success rate of 0.791.

Since the benchmark functions Rastrigin and Alpine are more challenging, we use more particles N and a larger number of steps K , namely $K = 200$ and $K = 500$. We report the results in Table 3.

Number of steps $K = 200$						
Test function	M	$N = 300$	$N = 600$	$N = 900$	$N = 1200$	$N = 1500$
Rastrigin	1	0.180	0.256	0.298	0.322	0.337
	$+\infty$	0	0	0.004	0.004	0.007
Alpine	1	0.029	0.049	0.051	0.070	0.080
	$+\infty$	0	0.001	0.004	0.004	0.004
Number of steps $K = 500$						
Test function	M	$N = 300$	$N = 600$	$N = 900$	$N = 1200$	$N = 1500$
Rastrigin	1	0.213	0.265	0.316	0.326	0.343
	$+\infty$	0.001	0.004	0.005	0.009	0.010
Alpine	1	0.103	0.115	0.147	0.165	0.173
	$+\infty$	0.010	0.015	0.033	0.037	0.040

Table 3: For the 15-dimensional Rastrigin and Alpine function, both algorithms have difficulties in finding the global minimizer. However, the success rates for the CBO method with truncation ($M = 1$) are significantly higher compared to those of the original CBO method ($M = +\infty$).

Anisotropic Case. Let $d = 20$. In the case of anisotropic noise, we set $\lambda = 1$, $\sigma = 5$, $\alpha = 10^5$ and step-size $\Delta t = 0.02$. The initial positions of the particles are initialized with $\rho_0 = \mathcal{N}(0, 100I_d)$.

In Table 4 we report results comparing the anisotropic CBO method with truncation $M = 1$ and the original anisotropic CBO method [13, 24] ($M = +\infty$) for the Rastrigin, Ackley, Griewank and Salomon function. Each algorithm is run for $K = 200$ steps.

Number of steps $K = 1000$						
Test function	M	$N = 75$	$N = 150$	$N = 300$	$N = 600$	$N = 900$
Rastrigin	1	0.285	0.928	0.990	1	1
	$+\infty$	0.728	0.952	0.993	1	1
Ackley	1	0.510	0.997	1	1	1
	$+\infty$	0.997	1	1	1	1
Griewank	1	0.097	0.458	0.576	0.625	0.665
	$+\infty$	0.093	0.101	0.157	0.159	0.167
Salomon	1	0.010	0.434	0.925	0.998	1
	$+\infty$	0.622	0.954	0.970	0.934	0.891

Table 4: For the 20-dimensional Rastrigin, Ackley and Salomon function, the original anisotropic CBO method ($M = +\infty$) works better than the anisotropic CBO method with truncation ($M = 1$), in particular when the particle number N is small. In the case of the Salomon function, when increasing the number of particle to $N = 900$, the success rates of the original anisotropic CBO method ($M = +\infty$) decreases. In the case of the Griewank function, however, we find that the anisotropic CBO method with truncation ($M = +\infty$) works considerably better than the original anisotropic CBO method ($M = 1$).

Since the benchmark function Alpine is more challenging and none of the algorithms work in the previous setting, we reduce the dimensionality to $d = 15$, choose $\sigma = 1$, use $\rho_0 = \mathcal{N}(0, I_d)$ to initialize, employ more particles and use a larger number of steps K , namely $K = 200$, $K = 500$ and $K = 1000$. We report the results in Table 5.

Number of steps $K = 200$						
Test function	M	$N = 300$	$N = 600$	$N = 900$	$N = 1200$	$N = 1500$
Alpine	1	0	0.006	0.006	0.008	0.025
	$+\infty$	0.001	0.004	0.008	0.007	0.021
Number of steps $K = 500$						
Test function	M	$N = 300$	$N = 600$	$N = 900$	$N = 1200$	$N = 1500$
Alpine	1	0.130	0.224	0.291	0.336	0.365
	$+\infty$	0.083	0.175	0.250	0.292	0.330
Number of steps $K = 1000$						
Test function	M	$N = 300$	$N = 600$	$N = 900$	$N = 1200$	$N = 1500$
Alpine	1	0.102	0.198	0.293	0.340	0.368
	$+\infty$	0.097	0.179	0.250	0.295	0.331

Table 5: For the 15-dimensional Alpine function, the anisotropic CBO method with truncated noise ($M = 1$) works better than the original anisotropic CBO method ($M = +\infty$).

5 Conclusions

In this paper we establish the convergence to a global minimizer of a potentially nonconvex and nonsmooth objective function for a variant of consensus-based optimization (CBO) which incorporates truncated noise. We observe that truncating the noise in CBO enhances the well-behavedness of the statistics of the law of the dynamics, which enables enhanced convergence

performance and allows in particular for a wider flexibility in choosing the noise parameter of the method, as we observe numerically. For rigorously proving the convergence of the implementable algorithm to the global minimizer of the objective, we follow the route devised in [23].

Acknowledgements and Competing Interests

This work has been funded by the KAUST Baseline Research Scheme and the German Federal Ministry of Education and Research, and the Bavarian State Ministry for Science and the Arts. In addition to this, MF acknowledges the support of the Munich Center for Machine Learning. PR acknowledges the support of the Extreme Computing Research Center at KAUST. KR acknowledges the support of the Munich Center for Machine Learning and the financial support from the Technical University of Munich – Institute for Ethics in Artificial Intelligence (IEAI). LS acknowledges the support of KAUST Optimization and Machine Learning Lab. LS also thanks the hospitality of the Chair of Applied Numerical Analysis of the Technical University of Munich for discussions that contributed to the finalization of this work.

References

- [1] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines. A stochastic approach to combinatorial optimization and neural computing*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Ltd., Chichester, 1989.
- [2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
- [3] T. Bäck, D. B. Fogel, and Z. Michalewicz, editors. *Handbook of evolutionary computation*. Institute of Physics Publishing, Bristol; Oxford University Press, New York, 1997.
- [4] H.-O. Bae, S.-Y. Ha, M. Kang, H. Lim, C. Min, and J. Yoo. A constrained consensus based optimization algorithm and its application to finance. *Appl. Math. Comput.*, 416:Paper No. 126726, 10, 2022.
- [5] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*, 35(3):268–308, 2003.
- [6] G. Borghi, S. Grassi, and L. Pareschi. Consensus based optimization with memory effects: random selection and applications. *arXiv preprint arXiv:2301.13242*, 2023.
- [7] G. Borghi, M. Herty, and L. Pareschi. A consensus-based algorithm for multi-objective optimization and its mean-field description. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4131–4136. IEEE, 2022.
- [8] G. Borghi, M. Herty, and L. Pareschi. An adaptive consensus based method for multi-objective optimization with uniform Pareto front approximation. *Applied Mathematics & Optimization*, 88(2):1–43, 2023.
- [9] G. Borghi, M. Herty, and L. Pareschi. Constrained consensus-based optimization. *SIAM Journal on Optimization*, 33(1):211–236, 2023.
- [10] L. Bungert, P. Wacker, and T. Roith. Polarized consensus-based dynamics for optimization and sampling. *arXiv:2211.05238*, 2022.

- [11] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28(6):1037–1066, 2018.
- [12] J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes. Consensus-based sampling. *Stud. Appl. Math.*, 148(3):1069–1140, 2022.
- [13] J. A. Carrillo, S. Jin, L. Li, and Y. Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.*, 27(suppl.):Paper No. S5, 22, 2021.
- [14] J. A. Carrillo, C. Totzeck, and U. Vaes. Consensus-based optimization and ensemble kalman inversion for global optimization problems with constraints. In *Modeling and Simulation for Collective Dynamics*, pages 195–230. World Scientific, 2023.
- [15] J. A. Carrillo, N. G. Trillos, S. Li, and Y. Zhu. FedCBO: Reaching group consensus in clustered federated learning through consensus-based optimization. *arXiv:2305.02894*, 2023.
- [16] J. Chen, S. Jin, and L. Lyu. A consensus-based global optimization method with adaptive momentum estimation. *arXiv:2012.04827*, 2020.
- [17] C. Cipriani, H. Huang, and J. Qiu. Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Appl. Math.*, 54(3):3091–3121, 2022.
- [18] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.
- [19] D. B. Fogel. *Evolutionary computation. Toward a new philosophy of machine intelligence*. IEEE Press, Piscataway, NJ, second edition, 2000.
- [20] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.*, 30(14):2725–2751, 2020.
- [21] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.*, 22:Paper No. 237, 55, 2021.
- [22] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Anisotropic diffusion in consensus-based optimization on the sphere. *SIAM J. Optim.*, 32(3):1984–2012, 2022.
- [23] M. Fornasier, T. Klock, and K. Riedl. Consensus-based optimization methods converge globally. *arXiv:2103.15130*, 2021.
- [24] M. Fornasier, T. Klock, and K. Riedl. Convergence of anisotropic consensus-based optimization in mean-field law. In J. L. Jiménez Laredo, J. I. Hidalgo, and K. O. Babaagba, editors, *Applications of Evolutionary Computation*, pages 738–754, Cham, 2022. Springer International Publishing.
- [25] S. Grassi, H. Huang, L. Pareschi, and J. Qiu. Mean-field particle swarm optimization. In *Modeling and Simulation for Collective Dynamics*, pages 127–193. World Scientific, 2023.
- [26] S. Grassi and L. Pareschi. From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Math. Models Methods Appl. Sci.*, 31(8):1625–1657, 2021.

- [27] S.-Y. Ha, S. Jin, and D. Kim. Convergence of a first-order consensus-based global optimization algorithm. *Math. Models Methods Appl. Sci.*, 30(12):2417–2444, 2020.
- [28] S.-Y. Ha, S. Jin, and D. Kim. Convergence and error estimates for time-discrete consensus-based optimization algorithms. *Numer. Math.*, 147(2):255–282, 2021.
- [29] S.-Y. Ha, M. Kang, and D. Kim. Emergent behaviors of high-dimensional Kuramoto models on Stiefel manifolds. *Automatica*, 136:Paper No. 110072, 2022.
- [30] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001.
- [31] J. H. Holland. *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor, Mich., 1975.
- [32] H. Huang and J. Qiu. On the mean-field limit for the consensus-based optimization. *Math. Methods Appl. Sci.*, 45(12):7814–7831, 2022.
- [33] H. Huang, J. Qiu, and K. Riedl. Consensus-based optimization for saddle point problems. *arXiv:2212.12334*, 2022.
- [34] H. Huang, J. Qiu, and K. Riedl. On the global convergence of particle swarm optimization methods. *Applied Mathematics & Optimization*, 88(2):30, 2023.
- [35] D. Kalise, A. Sharma, and M. V. Tretyakov. Consensus-based optimization via jump-diffusion stochastic differential equations. *Mathematical Models and Methods in Applied Sciences*, 33(02):289–339, 2023.
- [36] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN’95 - International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.
- [37] J. Kim, M. Kang, D. Kim, S.-Y. Ha, and I. Yang. A stochastic consensus method for nonconvex optimization on the Stiefel manifold. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1050–1057. IEEE, 2020.
- [38] K. Klamroth, M. Stiglmayr, and C. Totzeck. Consensus-based optimization for multi-objective problems: A multi-swarm approach. *arXiv:2211.15737*, 2022.
- [39] D. Ko, S.-Y. Ha, S. Jin, and D. Kim. Convergence analysis of the discrete consensus-based optimization algorithm with random batch interactions and heterogeneous noises. *Mathematical Models and Methods in Applied Sciences*, 32(06):1071–1107, 2022.
- [40] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.
- [41] E. Platen. An introduction to numerical methods for stochastic differential equations. In *Acta numerica, 1999*, volume 8 of *Acta Numer.*, pages 197–246. Cambridge Univ. Press, Cambridge, 1999.
- [42] K. Riedl. Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law. *European Journal of Applied Mathematics*, First View:1–32, 2023.

- [43] K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier. Gradient is all you need? *arXiv:2306.09778*, 2023.
- [44] C. Schillings, C. Totzeck, and P. Wacker. Ensemble-based gradient inference for particle methods in optimization and sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):757–787, 2023.
- [45] C. Totzeck. Trends in consensus-based optimization. In *Active Particles, Volume 3: Advances in Theory, Models, and Applications*, pages 201–226. Springer, 2021.
- [46] C. Totzeck and M.-T. Wolfram. Consensus-based global optimization with personal best. *Math. Biosci. Eng.*, 17(5):6026–6044, 2020.

License for [CBO-III].

The permission to reprint and include the material is printed on the next page(s).

Von: EJAM ejam@cambridge.org
Betreff: RE: Request for Permission to use Material in my Dissertation (EJAM-D-23-00126)
Datum: 30. März 2024 um 21:46
An: Konstantin Riedl konstantin.riedl@ma.tum.de

Hello again,

Here is the answer ahead of the publication of your paper.

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

Best wishes,
Susie
Susie Cox
Editorial office
Cambridge University Press

Von: Konstantin Riedl konstantin.riedl@ma.tum.de
Betreff: Request for Permission to use Material in my Dissertation (EJAM-D-23-00126)
Datum: 14. März 2024 um 18:02
An: ejam@cambridge.org

Dear Sir or Madam,

As one of the authors of the article „Consensus-Based Optimization with Truncated Noise“ (EJAM-D-23-00126) which is accepted and currently in production for publication in the special issue *From integro-differential models to data-oriented approaches for emergent phenomena* of the *European Journal of Applied Mathematics*, I am reaching out to you to request permission to include the paper in my dissertation (doctoral thesis).

Since I am pursuing a cumulative dissertation, it is necessary by the rules of my university, the Technical University of Munich, and the School of Computation, Information and Technology, that I provide and include in my dissertation a **written letter of approval from the publisher** for all my publications that are part of my dissertation.

I hereby kindly ask for such a confirmation from Cambridge University Press (by email or a weblink to your terms and conditions), which allows me to use the aforementioned article in my dissertation.

If you have any questions beforehand, do not hesitate to ask.

Best regards,
Konstantin


Konstantin Riedl, M.Sc.

Technical University of Munich
School of Computation, Information and Technology
Department of Mathematics
Chair for Applied Numerical Analysis

Munich Center for Machine Learning

Institute for Ethics in Artificial Intelligence

Email: konstantin.riedl@ma.tum.de

Products and Services  Discover Content 

Cambridge Core

[Home](#) > [Journals](#) > [European Journal of Applied Mathematics](#) > [Journal information](#) > [Journal policies](#)
> [Open access options](#)

English | [Français](#)

European Journal of Applied Mathematics

Search within full text

Other actions



Open access options

This journal is a wholly [open access](https://www.cambridge.org/core/services/open-access-policies/introduction-to-open-access) journal, which means all articles are published as **Gold Open Access** under a Creative Commons licence. This enables anyone to access and redistribute the content and, depending upon the licence, re-use the content in new or derivative works with attribution. The terms of re-use for Gold Open Access content are stated in the copyright line of the article.

Authors may choose any of the six [Creative Commons licences](https://www.cambridge.org/core/services/open-access-policies/open-access-resources/creative-commons-licenses) when publishing Gold Open Access in this journal.

Open Access Funding

Multiple funding routes for Gold Open Access publication are available to authors publishing in this journal:

- If the corresponding author of a research article is affiliated with an institution with which Cambridge University Press has an [open access publishing agreement](#), in the majority of cases this will cover the full costs of publication (specific terms and conditions vary) and Cambridge will liaise directly with the institution to determine eligibility and secure this funding. Authors can use our [eligibility checker](#) to see if their institution has an active agreement.
- The [Cambridge Open Equity Initiative](#) (COEI) funds Gold Open Access publishing for authors from over 100 low- and middle-income countries, covering over 5,000 institutions. Eligibility is automatically established during the publication process and no Gold Open Access fees will be charged if the corresponding author of a research article is based in one of these countries.
- If the research behind the article was funded by a grant or arrangement with a funding body that supports the payment of article processing charges (APCs), authors are asked to pay an APC out of those funds. Please check the [fees and pricing](#) page for details of this journal's charges, and our [central APC page](#) for further information such as APC refund policies.
- Authors from low- and middle-income countries who are not covered by an institutional open access agreement or the Cambridge Open Equity Initiative may be eligible for support under the [Research4Life](#) scheme. Corresponding authors of research articles based in 'Group A' countries will automatically have APC costs waived entirely, and authors based in 'Group B' countries will automatically receive a 50% discount on an article's APC.
- This journal also grants waivers on a discretionary basis, for authors who do not have funds available to pay an APC and are not covered by the options above. Authors must request a waiver at or before submission, before an article enters editorial consideration, by filling out a [waiver request form](#).

Please note that the decision whether to accept a paper for publication will rest solely with the Editor, and without reference to the funding situation of the authors. The Editor, editorial board members, and reviewers will have no involvement with the billing of APCs and cannot grant discretionary waivers.

Other routes to open access

Under this journal's **Green Open Access** policy, authors can make pre-published versions of their articles available in institutional or other repositories, or on their personal or departmental websites, under certain conditions. This allows authors to comply with the open access mandates of many funders and institutions before the final article is published as Gold Open Access. For more information, please see Cambridge University Press's [Green Open Access \(https://www.cambridge.org/core/services/open-access-policies/open-access-journals/green-open-access-policy-for-journals\)](https://www.cambridge.org/core/services/open-access-policies/open-access-journals/green-open-access-policy-for-journals) policy page, where you can download a spreadsheet with full details about which versions of articles authors can post online, and where and when authors can post them.

If you have open access questions which are not answered by our [policy pages and resources \(https://www.cambridge.org/openaccess\)](https://www.cambridge.org/openaccess), please contact [openresearch@cambridge.org \(mailto:openresearch@cambridge.org\)](mailto:openresearch@cambridge.org).

Social Sharing

This journal participates in [Cambridge Core Share \(https://www.cambridge.org/coreshare\)](https://www.cambridge.org/coreshare), an initiative that allows a read-only version of a final published PDF (the Version of Record) to be shared and easily accessed by anyone. Core Share links, and Core Share PDFs containing the links, can be freely shared on social media sites and scholarly collaboration networks to enhance both the impact and discoverability of research.

Preprint policy

A preprint is an early version of an article prior to the version accepted for publication in a journal. This journal allows preprints to be posted anywhere at any time, including before submission to the journal. For more information, see the Cambridge University Press [preprint policy \(https://www.cambridge.org/core/services/open-access-policies/open-access-journals/preprint-policy\)](https://www.cambridge.org/core/services/open-access-policies/open-access-journals/preprint-policy).



Products and Services ▼

 Discover Content ▼

Cambridge Core

[Home](#) > [Services](#) > [Open access policies](#) > [Open access journals](#) > [Preprint policy](#)

Preprint policy

Preprint policy at Cambridge University Press

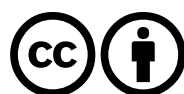
A 'preprint' is an early version of a manuscript, created prior to the version accepted for publication.

Our editorial guidelines require that manuscripts submitted to Cambridge University Press not be previously published, or be under simultaneous review for publication, in part or in whole within an academic journal, book or similar entity. However, deposition of a preprint anywhere else shall not be viewed as prior publication.

Upon acceptance of a manuscript for publication the author must agree to the terms of the relevant publishing agreement. After the final version of the work is published, the preprint can still be shared and used under its original licence terms.

It is best practice to link preprints and the final published work (the 'Version of Record'). If appropriate, we encourage authors to include details of preprint posting, including DOI or other persistent identifier, when submitting their manuscript. Authors are also encouraged to ensure that the preprint record is later updated with a DOI and a URL link to the Version of Record, if their manuscript is accepted and published by Cambridge University Press.

For further information, please see our Green Open Access policies for [journals](https://www.cambridge.org/core/services/open-access-policies/open-access-journals/green-open-access-policy-for-journals) (<https://www.cambridge.org/core/services/open-access-policies/open-access-journals/green-open-access-policy-for-journals>), [books](https://www.cambridge.org/core/services/open-access-policies/open-access-books/green-open-access-policy-for-books) (<https://www.cambridge.org/core/services/open-access-policies/open-access-books/green-open-access-policy-for-books>), and [Elements](https://www.cambridge.org/core/services/open-access-policies/open-access-elements/green-open-access-policy-for-elements) (<https://www.cambridge.org/core/services/open-access-policies/open-access-elements/green-open-access-policy-for-elements>), or contact openresearch@cambridge.org (<mailto:openresearch@cambridge.org>).



CC BY 4.0 DEED

Attribution 4.0 International

Canonical URL : <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict

others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation** .

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- [Learn more about our work](#)
- **[Learn more about CC Licensing](#)**
- [Support our work](#)
- [Use the license for your own material.](#)
- [Licenses List](#)
- [Public Domain List](#)

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- [More info](#)

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indication of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- [Marking guide](#)
- [More info](#)

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WIPO Copyright Treaty.

- [More info](#)

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- [More info](#)

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- [More info](#)

[Contact](#)

[Newsletter](#)

[Privacy](#)

[Policies](#)

[Terms](#)

CONTACT US

Creative Commons PO Box 1866,
Mountain View, CA 94042

info@creativecommons.org

+1 415 429 6753

SUBSCRIBE TO OUR NEWSLETTER

SUBSCRIBE

SUPPORT OUR WORK

Our work relies on you! Help us keep the Internet free and open.

DONATE NOW

Except where otherwise **noted**, content on this site is licensed under a **Creative Commons Attribution 4.0 International license**. Icons by **Font Awesome**.

Paper P4

Leveraging Memory Effects and
Gradient Information in
Consensus-Based Optimization:
On Global Convergence
in Mean-Field Law

K. Riedl

Eur. J. Appl. Math. (accepted 2023, to appear)

Paper Summary of [CBO-IV]³⁷

In the paper “Leveraging Memory Effects and Gradient Information in Consensus-Based Optimisation: On Global Convergence in Mean-Field Law,” published in the *European Journal of Applied Mathematics*, we propose and, by taking a mean-field perspective, theoretically analyze the variant (3.117) of CBO, which makes use of memory effects and gradient information.

CBO is a versatile, flexible and customizable multi-particle metaheuristic optimization method suitable for performing nonconvex and nonsmooth global optimizations in the form of (2.1) in high dimensions. It has proven effective and successful in various applications, while remaining at the same time amenable to theoretical analysis thanks to its originally minimalistic design.

The underlying dynamics, however, is flexible enough to incorporate different mechanisms widely used in evolutionary computation and machine learning, as we demonstrate in [CBO-IV] by analyzing a variant of CBO, given in its discrete-time form in (3.117), which exploits memory mechanisms and gradient information. For modeling the memory of a particle, we follow the work [GP21]. A generalization of the analytical techniques put forward in [CBO-I; CBO-II] then allows us to rigorously prove that this more elaborate dynamics converges to a global minimizer of the objective function in mean-field law for a vast class of functions under minimal assumptions on the initialization of the method [CBO-IV, Theorem 5]. The proof, in particular, reveals how to leverage further, in some applications advantageous, forces in the dynamics without losing provable global convergence in the mean-field limit. A study of the mean-field approximation of this variant is left for future considerations. As a corollary we present an analogous convergence result for the CBO dynamics with gradient information but without memory effects [CBO-IV, Corollary 2.6]. To demonstrate the benefit of the herein investigated memory effects and gradient information in certain applications, we present numerical evidence for the superiority of this CBO variant for benchmark problems in optimization, see [CBO-IV, Figure 2], as well as in applications coming from machine learning and compressed sensing [CBO-IV, Section 4]. More specifically, by reusing the efficient implementation of the anisotropic CBO algorithm from [CBO-II], which utilizes several tweaks in the implementation, such as random mini-batch ideas and a cooling strategy of the parameters [Car+21; For+21], and which is now enhanced through the use of memory mechanics, we train both a shallow and a convolutional neural network model for classifying the MNIST dataset of handwritten digits [LCB10]. Moreover, by approaching an experiment in compressed sensing [FR13], which has become a very active and profitable field of research since the seminal works [CRT06; Don06], we showcase an application of CBO methods, where gradient information turns out to be indispensable for their success.

KR’s Contributions. KR is the sole author of this work.

³⁷In this section, we follow [CBO-IV, Abstract].

The following document is a reprint of

[CBO-IV] K. Riedl. “Leveraging Memory Effects and Gradient Information in Consensus-Based Optimisation: On Global Convergence in Mean-Field Law.” In: *Eur. J. Appl. Math.* (accepted 2023, to appear), 32 pages.

The permission to reprint and include the material is provided after the reprint.

PAPER

Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law

Konstantin Riedl^{1,2} 

¹Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, Munich, Germany and ²Munich Center for Machine Learning, Munich, Germany

Email: konstantin.riedl@ma.tum.de

Received: 23 November 2022; **Revised:** 02 August 2023; **Accepted:** 15 September 2023

Keywords: High-dimensional global optimisation; metaheuristics; consensus-based optimisation; mean-field limit; Fokker-Planck equations

2020 Mathematics Subject Classification: 65K10 (Primary); 90C26, 90C56, 35Q90, 35Q84 (Secondary)

Abstract

In this paper, we study consensus-based optimisation (CBO), a versatile, flexible and customisable optimisation method suitable for performing nonconvex and nonsmooth global optimisations in high dimensions. CBO is a multi-particle metaheuristic, which is effective in various applications and at the same time amenable to theoretical analysis thanks to its minimalistic design. The underlying dynamics, however, is flexible enough to incorporate different mechanisms widely used in evolutionary computation and machine learning, as we show by analysing a variant of CBO which makes use of memory effects and gradient information. We rigorously prove that this dynamics converges to a global minimiser of the objective function in mean-field law for a vast class of functions under minimal assumptions on the initialisation of the method. The proof in particular reveals how to leverage further, in some applications advantageous, forces in the dynamics without losing provable global convergence. To demonstrate the benefit of the herein investigated memory effects and gradient information in certain applications, we present numerical evidence for the superiority of this CBO variant in applications such as machine learning and compressed sensing, which en passant widen the scope of applications of CBO.

1. Introduction

Interacting multi-particle systems are ubiquitous in a wide variety of scientific disciplines with application areas reaching from atomic scales over the human scale to the astronomical scale. For instance, large-scale multi-agent models are used to understand the coordinated movement of animal groups [19, 51] or crowds of people [1, 20]. Especially fascinating in this context is that such complex and often intelligent behaviour – phenomena known as self-organisation and swarm intelligence – emerge from seemingly simple rules of interaction [56]. This intriguing capabilities have drawn researchers' attention towards designing interacting particle systems for specific purposes in various disciplines. In applied mathematics in particular, agent-based optimisation algorithms look back on a long and successful history of empirically achieving state-of-the-art performance on challenging global unconstrained problems of the form

$$x^* = \arg \min_{x \in \mathbb{R}^d} \mathcal{E}(x).$$

Here, $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ denotes a possibly nonconvex and nonsmooth high-dimensional objective function, whose global minimiser x^* is assumed to exist and be unique for the remainder of this work. Well-known representatives of this family are Evolutionary Programming [24], Genetic Algorithms [38], Particle



Swarm Optimisation [43] and Ant Colony Optimisation [23]. They belong to the broad class of so-called metaheuristics [4, 6], which are methods orchestrating an interaction between local improvement procedures and global strategies, deterministic and stochastic processes, to eventually design an efficient and robust procedure for searching the solution space of the objective function \mathcal{E} .

Motivated by both the substantiated success of metaheuristics in applications and the lack of rigorous theoretical guarantees about their convergence and performance, the authors of [52] proposed consensus-based optimisation (CBO), which follows the spirit of metaheuristics but allows for a rigorous theoretical analysis [12, 14, 28, 29, 33, 34]. By taking inspiration from consensus formation in opinion dynamics [36], CBO methods use N particles X^1, \dots, X^N to explore the energy landscape of the objective \mathcal{E} and to eventually form a consensus about the location of the global minimiser x^* . In its original form [52], the dynamics of each particle X^i , which is governed by a stochastic differential equation (SDE), is subject to two competing forces. A deterministic drift term pulls the particles towards a so-called consensus point, which is an instantaneously computed weighted average of the positions of all particles and approximates the global minimiser x^* the best possible given the currently available information. The resulting contractive behaviour is counteracted by the second term which is stochastic in nature and thereby features the exploration of the energy landscape of the objective function. Its magnitude and therefore its explorative power scales with the distance of the individual particle from the consensus point, which encourages particles far away to explore larger regions of the domain, while particles already close advance their position only locally.

In this work, motivated by the numerical evidence presented below as well as other recent papers such as [32, 54, 55], we consider a more elaborate variant of this dynamics which exhibits the two following additional drift terms.

- The first is a drift towards the historical best position of the particular particle. To store such information, we follow the work [32], where the authors introduce for each particle an additional state variable Y^i , which can be regarded as the memory of the respective particle X^i . In contrast to the original dynamics, an individual particle is therefore described by the tuple (X^i, Y^i) . Moreover, the consensus point is no longer computed from the instantaneous positions X^i , but the historical best positions Y^i .
- The second term is a drift in the direction of the negative gradient of \mathcal{E} evaluated at the current position of the respective particle X^i .

Both terms are accompanied by associated noise terms. We now make the CBO dynamics with memory effects and gradient information rigorous by providing a formal description of the interacting particle system. A visualisation of the dynamics with all relevant quantities and forces is provided in Figure 1. Given a finite time horizon $T > 0$, and user-specified parameters $\alpha, \beta, \theta, \kappa, \lambda_1, \sigma_1 > 0$ and $\lambda_2, \lambda_3, \sigma_2, \sigma_3 \geq 0$, the dynamics is given by the system of SDEs

$$dX_t^i = -\lambda_1(X_t^i - y_\alpha(\widehat{\rho}_{Y,t}^N)) dt - \lambda_2(X_t^i - Y_t^i) dt - \lambda_3 \nabla \mathcal{E}(X_t^i) dt + \sigma_1 D(X_t^i - y_\alpha(\widehat{\rho}_{Y,t}^N)) dB_t^{1,i} + \sigma_2 D(X_t^i - Y_t^i) dB_t^{2,i} + \sigma_3 D(\nabla \mathcal{E}(X_t^i)) dB_t^{3,i}, \quad (1.1a)$$

$$dY_t^i = \kappa (X_t^i - Y_t^i) S^{\beta,\theta}(X_t^i, Y_t^i) dt \quad (1.1b)$$

for $i = 1, \dots, N$ and where $((B_t^{m,i})_{t \geq 0})_{i=1, \dots, N}$ are independent standard Brownian motions in \mathbb{R}^d for $m \in \{1, 2, 3\}$. The system is complemented with independent initial data $(X_0^i, Y_0^i)_{i=1, \dots, N}$, typically such that $X_0^i = Y_0^i$ for all $i = 1, \dots, N$. A numerical implementation of the scheme usually originates from an Euler-Maruyama time discretisation of equation (1.1). The first term appearing in the SDE for the position X_t^i , i.e., in the first line of equation (1.1a), is the drift towards the consensus point

$$y_\alpha(\widehat{\rho}_{Y,t}^N) := \int y \frac{\omega_\alpha(y)}{\|\omega_\alpha\|_{L_1(\widehat{\rho}_{Y,t}^N)}} d\widehat{\rho}_{Y,t}^N(y), \quad \text{with } \omega_\alpha(y) := \exp(-\alpha \mathcal{E}(y)). \quad (1.2)$$

Here, $\widehat{\rho}_{Y,t}^N$ denotes the random empirical measure of the particles' historical best positions, i.e., $\widehat{\rho}_{Y,t}^N := \frac{1}{N} \sum_{i=1}^N \delta_{Y_t^i}$. Definition (1.2) is motivated by the fact that $y_\alpha(\widehat{\rho}_{Y,t}^N) \approx \arg \min_{i \in \{1, \dots, N\}} \mathcal{E}(Y_t^i)$ as $\alpha \rightarrow \infty$ under

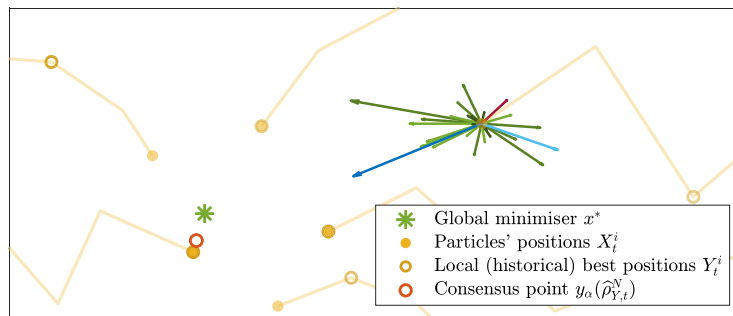


Figure 1. A visualisation of the CBO dynamics (1.1) with memory effects and gradient information. Particles with positions X^1, \dots, X^N (yellow dots with their trajectories) explore the energy landscape of the objective \mathcal{E} in search of the global minimiser x^* (green star). Each particle stores its local historical best position Y_t^i (yellow circles). The dynamics of the position X_t^i of each particle is governed by three deterministic terms with associated random noise terms (visualised by depicting eight possible realisations with differently shaded green arrows). A global drift term (dark blue arrow) drags the particle towards the consensus point $y_\alpha(\hat{\rho}_{Y,t}^N)$ (orange circle), which is computed as a weighted (visualised through colour opacity) average of the particles' historical best positions. A local drift term (light blue arrow) imposes movement towards the respective local best position Y_t^i . A gradient drift term (purple arrow) exerts a force in the direction $-\nabla\mathcal{E}(X_t^i)$.

reasonable assumptions. The first term in the second line of equation (1.1a) is with the consensus drift associated diffusion term, which injects randomness into the dynamics and thereby features the explorative nature of the algorithm. The two commonly studied diffusion types are isotropic [12, 28, 52] and anisotropic [14, 29] diffusion with

$$D(\cdot) = \begin{cases} \|\cdot\|_2 \text{Id}, & \text{for isotropic diffusion,} \\ \text{diag}(\cdot), & \text{for anisotropic diffusion,} \end{cases} \tag{1.3}$$

where $\text{Id} \in \mathbb{R}^{d \times d}$ is the identity matrix and $\text{diag} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ the operator mapping a vector onto a diagonal matrix with the vector as its diagonal. Despite the potential of the dynamics getting trapped in affine subspaces, the coordinate-dependent scaling of anisotropic diffusion has proven to be beneficial for the performance of the method in high-dimensional applications by allowing for dimension-independent convergence rates [14, 29]. For this reason, we restrict our attention to the case of anisotropic noise in what follows. Nevertheless, theoretically similar results as the ones presented in this work can be obtained also for the isotropic case. The second term in the first line of equation (1.1a) is the drift towards the historical best position of the respective particle. In contrast to the global nature of the consensus drift, which incorporates information from all N particles, this term depends only on the past of the specific particle. To store such information about the history of each particle [32], an additional state variable Y^i is introduced for every particle, which evolves according to equation (1.1b), where

$$S^{\beta,\theta}(x, y) = \frac{1}{2} (1 + \theta + \tanh(\beta (\mathcal{E}(y) - \mathcal{E}(x)))) \tag{1.4}$$

is chosen throughout this article, which is an approximation to the Heaviside function $H(x, y) = \mathbb{1}_{\mathcal{E}(x) < \mathcal{E}(y)}$ as $\theta \rightarrow 0$ and $\beta \rightarrow \infty$. The variable Y_t^i can therefore be regarded as the memory of the i th particle, i.e., as the location of the in-time best-seen position of X^i up to time t . This can be understood most easily when discretising (1.1b) as

$$Y_{k+1}^i = Y_k^i + \Delta t \kappa (X_{k+1}^i - Y_k^i) S^{\beta,\theta}(X_{k+1}^i, Y_k^i)$$

and noting that with parameter choices $\kappa = 1/\Delta t$, $\theta = 0$ and $\beta \gg 1$ it holds $Y_{k+1}^i = X_{k+1}^i$ if $\mathcal{E}(X_{k+1}^i) < \mathcal{E}(Y_k^i)$ and $Y_{k+1}^i = Y_k^i$ else. The third term in the first line of equation (1.1a) is the drift in the direction

of the negative gradient of \mathcal{E} , which is a local and instantaneous contribution. The remaining two terms are noise terms, which are associated with the formerly described memory and gradient drifts.

A theoretical convergence analysis of CBO can be carried out either by directly investigating the microscopic system (1.1) or its numerical time discretisation, as promoted for instance in a simplified setting in the works [33, 34], or alternatively, as done for example in [12, 14, 26, 28, 29], by analysing the macroscopic behaviour of the particle density through a mean-field limit associated with (1.1). Formally, such mean-field limit is given by the self-consistent nonlinear and nonlocal SDE

$$d\bar{X}_t = -\lambda_1(\bar{X}_t - y_\alpha(\rho_{Y,t})) dt - \lambda_2(\bar{X}_t - \bar{Y}_t) dt - \lambda_3 \nabla \mathcal{E}(\bar{X}_t) dt + \sigma_1 D(\bar{X}_t - y_\alpha(\rho_{Y,t})) dB_t^1 + \sigma_2 D(\bar{X}_t - \bar{Y}_t) dB_t^2 + \sigma_3 D(\nabla \mathcal{E}(\bar{X}_t)) dB_t^3, \tag{1.5a}$$

$$d\bar{Y}_t = \kappa (\bar{X}_t - \bar{Y}_t) S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) dt, \tag{1.5b}$$

which is complemented with initial datum $(\bar{X}_0, \bar{Y}_0) \sim \rho_0$, and where $\rho_t = \rho(t) = \text{Law}(\bar{X}_t, \bar{Y}_t)$ with marginal law $\rho_{Y,t}$ of \bar{Y}_t given by $\rho_{Y,t} = \rho_Y(t, \cdot) = \int d\rho_t(\cdot, y)$. The measure $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ in particular weakly satisfies the Fokker-Planck equation

$$\partial_t \rho_t = \text{div}_x((\lambda_1(x - y_\alpha(\rho_{Y,t})) + \lambda_2(x - y) + \lambda_3 \nabla \mathcal{E}(x))\rho_t) + \text{div}_y((\kappa(y - x)S^{\beta,\theta}(x, y))\rho_t) + \frac{1}{2} \sum_{k=1}^d \partial_{x_k x_k}^2 \left((\sigma_1^2 D(x - y_\alpha(\rho_{Y,t}))_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(x))_{kk}^2) \rho_t \right), \tag{1.6}$$

see Definition 2.1. Working with the partial differential equation (PDE) (1.6) instead of the interacting particle system (1.1) typically permits to employ more powerful technical tools, which result in stronger and deterministic statements about the long-time behaviour of the average agent density ρ . This analysis approach is rigorously justified by the mean-field approximation, i.e., the fact that the empirical particle measure $\hat{\rho}_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{(X_t^i, Y_t^i)}$ converges in some sense to the mean-field law ρ_t as the number of particles N tends to infinity. For the original CBO dynamics, a qualitative result about convergence in distribution is provided in [39], which is based on a tightness argument in the path space. More precisely, the authors of that work show that the sequence $\{\hat{\rho}^N\}_{N \geq 2}$ of $\mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$ -valued random variables is tight, which permits to employ Prokhorov’s theorem to obtain, up to a subsequence, some limiting measure, which turns out to be deterministic and at the same time satisfy the associated mean-field PDE. A more desirable quantitative approximation result, on the other hand, can be established by proving propagation of chaos, i.e., by establishing for instance

$$\max_{i=1, \dots, N} \sup_{t \in [0, T]} \mathbb{E} \left\| (X_t^i, Y_t^i) - (\bar{X}_t^i, \bar{Y}_t^i) \right\|_2^2 \leq CN^{-1} \quad \text{as } N \rightarrow \infty,$$

where $(\bar{X}_t^i, \bar{Y}_t^i)$ denote N i.i.d. copies of the mean-field dynamics (1.5). For the original variant of unconstrained CBO, this was first done in [28, Section 3.3]. To keep the focus of this work on the long-time behaviour of the CBO variant (1.6), a rigorous analysis of the mean-field approximation is left for future considerations.

Before summarising the contributions of the present paper, let us put our work into context by providing a comprehensive literature overview about the history, developments and achievements of CBO.

Versatility and flexibility of CBO: a literature overview

Since its introduction in the work [52], CBO has gained a significant amount of attention from various research groups. This has led to a vast variety of different developments, of both theoretical and applied nature, as well as what concerns the mathematical modelling and numerical analysis of the method. By interpreting CBO as a stochastic relaxation of gradient descent, the recent work [53] even establishes a connection between the worlds of derivative-free and gradient-based optimisation.

A first rigorous but local convergence proof of the mean-field limit of CBO to global minimisers is provided for the cases of isotropic and anisotropic diffusion in [12, 14], respectively. By analysing

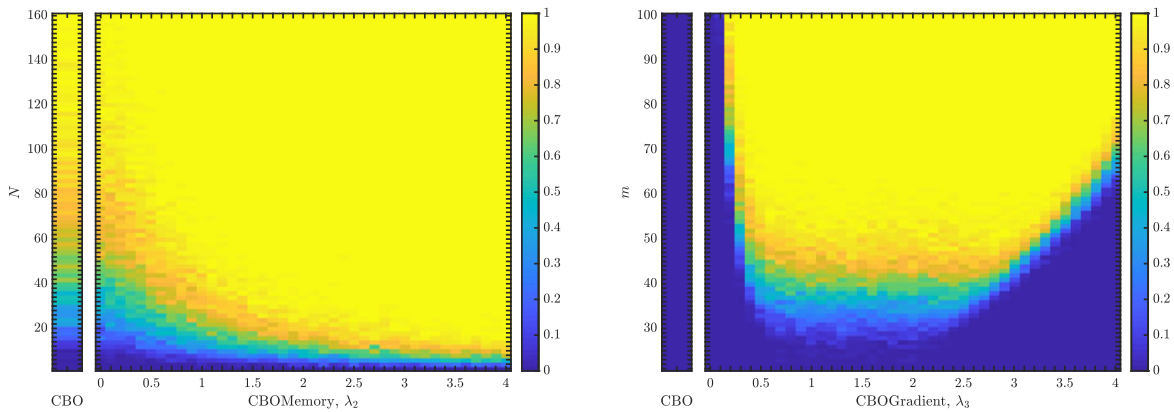
the time-evolution of the variance of the law of the mean-field dynamics ρ_t and proving its exponential decay towards zero, the authors first establish consensus formation at some stationary point before they ensure that this consensus is actually close to the global minimiser. A similarly flavoured approach is pursued in [33, 34], however, directly for the fully in-time discrete microscopic system and in the simplified setting, where the same Brownian motion is used for all agents, which limits the exploration capabilities of the method. In contrast, in the recent works [28, 29], the authors, again for the isotropic and anisotropic CBO variant, respectively, investigate the time-evolution of the Wasserstein-2 distance between the law ρ_t and a Dirac delta at the global minimiser. This is also the strategy which we pursue in this paper. By proving the exponential decay of $W_2(\rho_t, \delta_{x^*})$ to zero, consensus at the desired location follows immediately. Moreover, by providing a probabilistic quantitative result about the mean-field approximation, the authors give a first, and so far unique, holistic global convergence proof for the implementable, i.e., discretised numerical CBO algorithm in the unconstrained case. The results about the mean-field approximation of the latter papers were partially inspired by the series of works [25–27], in which the authors constrain the particle dynamics of CBO to compact hypersurfaces and prove local convergence of the numerical scheme to minimisers by adapting the technique of [12, 14]. This ensures a beneficial compactness of the stochastic processes, which simplifies the convergence of the interacting particle dynamics to the mean-field dynamics. In the unconstrained case, such intrinsic compactness is replaced by the fact that the dynamics are bounded with high probability, which is sufficient to establish convergence in probability. Further related works about CBO for optimisations with constraints include the papers [35, 44], where a problem on the Stiefel manifold is approached, and the works [9, 15], where the constrained optimisation is recast into a penalised problem. The philosophy of using an interacting swarm of particles to approach various relevant problems in science and engineering has promoted several variations of the original CBO algorithm for minimisation. Amongst them are methods based on consensus dynamics to tackle multi-objective optimisation problems [7, 8, 45], saddle point problems [40], the search for several minimisers simultaneously [10] or the sampling from certain distributions [13].

In the same vein and also in the spirit of this work, the original CBO method itself has undergone several modifications allowing for a more complex dynamics. This includes the use of particles with memory [31, 55], the integration of momentum [17], the usage of jump-diffusion processes [42] and the exploitation of on-the-fly extracted higher-order differential information through inferred gradients based on point evaluations of the objective function [54]. It moreover turned out that the renowned particle swarm optimisation method (PSO) [43] can be formulated and regarded as a second-order generalisation of CBO [18, 32]. This insight has enabled to adapt the for CBO-developed analysis techniques to rigorously prove the convergence of PSO [41].

In the collection of formerly referenced works and beyond, CBO has demonstrated to be a valuable method for a wide scope of applications reaching from the phase retrieval or robust subspace detection problem in signal processing [26, 27], over the training of neural networks for image classification in machine learning [14, 29] as well as in the setting of clustered federated learning [16], to asset allocation in finance [5]. It has been furthermore employed to approximate low-frequency functions in the presence of high-frequency noise and to the task of solving PDEs with low-regularity solutions [17].

Contributions

In view of the various developments and the wide scope of applications, a theoretical understanding of the long-time behaviour of the in practical applications employed CBO methods is of paramount interest. In this work, we analyse a variant of CBO which incorporates memory effects as well as gradient information from a theoretical and numerical perspective. As demonstrated concisely in Figure 2 and more comprehensively in Section 4, the herein investigated dynamics, which is more involved than standard CBO, proves to be beneficial in applications in machine learning and compressed sensing. Despite this additional complexity, by employing the analysis technique devised in [28, 29], we are able to provide



(a) Memory effects and an additional drift towards the historical best position of each individual particle improve the success probability of CBO.

(b) Gradient information and a drift in the direction of the negative gradient can be indispensable in certain applications such as compressed sensing.

Figure 2. A demonstration of the benefits of memory effects and gradient information in CBO methods. In both settings (a) and (b) the depicted success probabilities are averaged over 100 runs of CBO and the implemented scheme is given by a Euler-Maruyama discretisation of equation (1.1) with time horizon $T = 20$, discrete time step size $\Delta t = 0.01$, $\alpha = 100$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, $\lambda_1 = 1$ and $\sigma_1 = \sqrt{1.6}$. In (a) we plot the success probability of CBO without (left separate column) and with (right phase diagram) memory effects for different values of the parameter λ_2 , i.e., for different strengths of the memory drift, when optimising the Rastrigin function $\mathcal{E}(x) = \sum_{k=1}^d x_k^2 + \frac{5}{2}(1 - \cos(2\pi x_k))$ in dimension $d = 4$. As remaining parameters we choose $\sigma_2 = \lambda_1 \sigma_1$ and $\lambda_3 = \sigma_3 = 0$, i.e., no gradient information is involved. We observe that an increasing amount of memory drift improves the success probability significantly, even in the case where, theoretically, there are no convergence guarantees anymore, see Theorem 2.5 and Corollary 2.6. Section 4.2 provides further details. In (b) we depict the success probability of CBO without (left separate column) and with (right phase diagram) gradient information for different values of the parameter λ_3 , i.e., for different strengths of the gradient drift, when solving a compressed sensing problem in dimension $d = 200$ with sparsity $s = 8$. On the vertical axis we depict the number of measurements m , from which we try to recover the sparse signal by solving the associated ℓ_1 -regularised problem (LASSO). As remaining parameters we use merely $N = 10$ particles, choose $\sigma_3 = 0$ and $\lambda_2 = \sigma_2 = 0$, i.e., no memory drift is involved. We observe that gradient information is required to be able to identify the correct sparse solution and standard CBO would fail in such task. Section 4.4 provides more details.

rigorous mean-field-type convergence guarantees to the global minimiser, which describe the behaviour of the method in the large-particle limit and allow to draw conclusions about the typically observed performance in the practicable regime. Our results for CBO with memory effects and gradient information hold for a vast class of objective functions under minimal assumptions on the initialisation of the method. Moreover, the proof reveals how to leverage further, in other applications advantageous, forces in the dynamics while still being amenable to theory and allowing for provable global convergence.

1.1. Organisation

In Section 2, after providing details about the existence of solutions to the macroscopic SDE (1.5) and the associated PDE (1.6), we present and discuss our main theoretical contribution. It is about the convergence of CBO with memory effects and gradient information, as given in equation (1.1), to the global minimiser of the objective function in mean-field law, see [28, Definition 1]. More precisely, we show that the mean-field dynamics (1.5) and (1.6) converge with exponential rate to the global minimiser.

Section 3 contains the proof details of this result. In Section 4, we numerically demonstrate the benefits of the additional memory effects and gradient information of the previously analysed CBO variant. We in particular present applications of CBO in machine learning and compressed sensing, before we conclude the paper in Section 5.

For the sake of reproducible research, in the GitHub repository <https://github.com/KonstantinRiedl/CBOGlobalConvergenceAnalysis> we provide the Matlab code implementing the CBO algorithm with memory effects and gradient information analysed in this work.

1.2. Notation

Given a set $A \subset \mathbb{R}^d$, we write $(A)^c$ to denote its complement, i.e., $(A)^c := \{z \in \mathbb{R}^d : z \notin A\}$. For ℓ_∞ balls in \mathbb{R}^d with centre z and radius r , we write $B_r^\infty(z)$. The space of continuous functions $f: X \rightarrow Y$ is denoted by $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$ and a suitable topological space Y . For $X \subset \mathbb{R}^n$ open and for $Y = \mathbb{R}^m$, the function space $\mathcal{C}_c^k(X, Y)$ contains functions $f \in \mathcal{C}(X, Y)$ that are k -times continuously differentiable and have compact support. Y is omitted in the case of real-valued functions. The operator ∇ denotes the standard gradient of a function on \mathbb{R}^d .

In this paper, we mostly study laws of stochastic processes, $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, and we refer to a snapshot of such law at time t by writing $\rho_t \in \mathcal{P}(\mathbb{R}^d)$. Here, $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all Borel probability measures ϱ over \mathbb{R}^d . In $\mathcal{P}_p(\mathbb{R}^d)$ we moreover collect measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite p th moment. For any $1 \leq p < \infty$, W_p denotes the Wasserstein- p distance between two Borel probability measures $\varrho_1, \varrho_2 \in \mathcal{P}_p(\mathbb{R}^d)$, see, e.g., [2]. $\mathbb{E}(\varrho)$ denotes the expectation of a probability measure ϱ .

2. Global convergence in mean-field law

In the first part of this section, we provide an existence result about solutions of the nonlinear macroscopic SDE (1.5), respectively, the associated Fokker-Planck equation (1.6). Thereafter we specify the class of studied objective functions and present the main theoretical result about the convergence of the dynamics (1.5) and (1.6) to the global minimiser.

Throughout this work, we consider the – in typical applications beneficial – case of CBO with anisotropic diffusion, i.e., $D(\cdot) = \text{diag}(\cdot)$ in equations (1.1), (1.5) and (1.6), and also equation (2.1) below. However, up to minor modifications, analogous results can be obtained for isotropic diffusion.

2.1. Definition and existence of weak solutions

Let us begin by rigorously defining weak solutions of the Fokker-Planck equation (1.6).

Definition 2.1. Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, $T > 0$. We say $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ satisfies the Fokker-Planck equation (1.6) with initial condition ρ_0 in the weak sense in the time interval $[0, T]$, if we have for all $\phi \in \mathcal{C}_c^2(\mathbb{R}^d \times \mathbb{R}^d)$ and all $t \in (0, T)$

$$\begin{aligned} \frac{d}{dt} \int \int \phi(x, y) d\rho_t(x, y) &= - \int \int \kappa S^{\beta, \theta}(x, y) \langle y - x, \nabla_y \phi(x, y) \rangle d\rho_t(x, y) \\ &\quad - \int \int \lambda_1 \langle x - y_\alpha(\rho_{Y,t}), \nabla_x \phi(x, y) \rangle + \lambda_2 \langle x - y, \nabla_x \phi(x, y) \rangle + \lambda_3 \langle \nabla \mathcal{E}(x), \nabla_x \phi(x, y) \rangle d\rho_t(x, y) \\ &\quad + \frac{1}{2} \int \int \sum_{k=1}^d \left(\sigma_1^2 D(x - y_\alpha(\rho_{Y,t}))_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(x))_{kk}^2 \right) \partial_{x_k x_k}^2 \phi(x, y) d\rho_t(x, y) \end{aligned} \tag{2.1}$$

and $\lim_{t \rightarrow 0} \rho_t = \rho_0$ (in the sense of weak convergence of measures).

For solutions of the mean-field dynamics (1.5) and (1.6), we have the following existence result.

Theorem 2.2. Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$. Let $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$ with $\underline{\mathcal{E}} > -\infty$ satisfy for some constants $C_1, C_2 > 0$ the conditions

$$|\mathcal{E}(x) - \mathcal{E}(x')| \leq C_1 (\|x\|_2 + \|x'\|_2) \|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d, \tag{2.2}$$

$$\mathcal{E}(x) - \underline{\mathcal{E}} \leq C_2 (1 + \|x\|_2^2), \quad \text{for all } x \in \mathbb{R}^d, \tag{2.3}$$

and either $\sup_{x \in \mathbb{R}^d} \mathcal{E}(x) < \infty$ or

$$\mathcal{E}(x) - \underline{\mathcal{E}} \geq C_3 \|x\|_2^2, \quad \text{for all } \|x\|_2 \geq C_4 \tag{2.4}$$

for some $C_3, C_4 > 0$. Furthermore, in the case of an active gradient drift in the CBO dynamics (1.5), i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in C^1(\mathbb{R}^d)$ and obey additionally

$$\|\nabla \mathcal{E}(x) - \nabla \mathcal{E}(x')\|_2 \leq \tilde{L}_{\nabla \mathcal{E}} \|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d \tag{2.5}$$

for some $\tilde{L}_{\nabla \mathcal{E}} > 0$. Then, if (\bar{X}_0, \bar{Y}_0) is distributed according to ρ_0 , there exists a nonlinear process $(\bar{X}, \bar{Y}) \in \mathcal{C}([0, T], \mathbb{R}^d \times \mathbb{R}^d)$ satisfying (1.5) with associated law $\rho = \text{Law}((\bar{X}, \bar{Y}))$ having regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ and being a weak solution to the Fokker-Planck equation (1.6) with $\rho(0) = \rho_0$.

Assumption (2.2) requires that \mathcal{E} is locally Lipschitz continuous with the Lipschitz constant being allowed to have linear growth. This entails in particular that the objective has at most quadratic growth at infinity as formulated explicitly in Assumption (2.3), which can be seen when choosing $x' = x^*$ and $C_2 = 2C_1 \max\{1, \|x^*\|_2^2\}$ in (2.2). Assumption (2.4), on the other hand, assumes that \mathcal{E} also has at least quadratic growth in the farfield, i.e., overall it grows quadratically far away from x^* . Alternatively, \mathcal{E} may be bounded from above. Since the objective \mathcal{E} can be usually modified for the purpose of analysis outside a sufficiently large region, these growth conditions are not really restrictive. In case of an additional gradient drift term in the dynamics, i.e., $\lambda_3 \neq 0$, the objective naturally needs to be continuously differentiable. Furthermore, Assumption (2.5) imposes \mathcal{E} to be $\tilde{L}_{\nabla \mathcal{E}}$ -smooth, i.e., having an $\tilde{L}_{\nabla \mathcal{E}}$ -Lipschitz continuous gradient.

Remark 2.3. The regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ obtained in Theorem 2.2 above is an immediate consequence of the regularity of the initial condition $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$. It allows to extend the test function space $C_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ in Definition 2.1 to the larger space

$$C_*^2(\mathbb{R}^d \times \mathbb{R}^d) := \left\{ \phi \in C^2(\mathbb{R}^d \times \mathbb{R}^d) : \begin{aligned} &|\partial_{x_k} \phi(x, y)| \leq C_\phi (1 + \|x\|_2 + \|y\|_2) \text{ for some } C_\phi > 0 \\ &\text{and } \sup_{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d} \max_{k=1, \dots, d} |\partial_{x_k x_k}^2 \phi(x, y)| < \infty \end{aligned} \right\}, \tag{2.6}$$

as can be seen from the proof of Theorem 2.2, which we sketch in what follows.

Proof sketch of Theorem 2.2. The proof is based on the Leray-Schauder fixed point theorem and follows the steps taken in [12, Theorems 3.1, 3.2].

Step 1: For a given function $u \in \mathcal{C}([0, T], \mathbb{R}^d)$ and an initial measure $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$, according to standard SDE theory [3, Chapters 6], we can uniquely solve the auxiliary SDE

$$\begin{aligned} d\tilde{X}_t &= -\lambda_1(\tilde{X}_t - u_t) dt - \lambda_2(\tilde{X}_t - \tilde{Y}_t) dt - \lambda_3 \nabla \mathcal{E}(\tilde{X}_t) dt \\ &\quad + \sigma_1 D(\tilde{X}_t - u_t) dB_t^1 + \sigma_2 D(\tilde{X}_t - \tilde{Y}_t) dB_t^2 + \sigma_3 D(\nabla \mathcal{E}(\tilde{X}_t)) dB_t^3 \\ d\tilde{Y}_t &= \kappa (\tilde{X}_t - \tilde{Y}_t) S^{\beta, \theta}(\tilde{X}_t, \tilde{Y}_t) dt \end{aligned}$$

with $(\tilde{X}_0, \tilde{Y}_0) \sim \rho_0$. This is due to the fact that the coefficients of the drift and diffusion terms are locally Lipschitz continuous and have at most linear growth, which, in turn, is a consequence of the assumptions on \mathcal{E} as well as the smoothness of $S^{\beta, \theta}$ as defined in (1.4). This induces $\tilde{\rho}_t = \text{Law}((\tilde{X}_t, \tilde{Y}_t))$. Moreover, the assumed regularity of the initial distribution $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$ allows to obtain a fourth-order moment estimate of the form $\mathbb{E}[\|\tilde{X}_t\|_2^4 + \|\tilde{Y}_t\|_2^4] \leq (1 + 2\mathbb{E}[\|\tilde{X}_0\|_2^4 + \|\tilde{Y}_0\|_2^4])e^{ct}$, see, e.g. [3, Chapter 7]. So, in particular, $\tilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$.

Step 2: For some test function $\phi \in C_*^2(\mathbb{R}^d \times \mathbb{R}^d)$ as defined in (2.6), by Itô’s formula, we derive

$$\begin{aligned} d\phi(\tilde{X}_t, \tilde{Y}_t) &= \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (-\lambda_1(\tilde{X}_t - u_t) - \lambda_2(\tilde{X}_t - \tilde{Y}_t) - \lambda_3 \nabla \mathcal{E}(\tilde{X}_t)) dt \\ &\quad + \nabla_y \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (\kappa(\tilde{X}_t - \tilde{Y}_t) S^{\beta, \theta}(\tilde{X}_t, \tilde{Y}_t)) dt \\ &\quad + \frac{1}{2} \sum_{k=1}^d \partial_{x_k x_k}^2 \phi(\tilde{X}_t, \tilde{Y}_t) \left(\sigma_1^2 D(\tilde{X}_t - u_t)_{kk}^2 + \sigma_2^2 D(\tilde{X}_t - \tilde{Y}_t)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(\tilde{X}_t))_{kk}^2 \right) dt \\ &\quad + \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (\sigma_1 D(\tilde{X}_t - u_t) dB_t^1 + \sigma_2 D(\tilde{X}_t - \tilde{Y}_t) dB_t^2 + \sigma_3 D(\nabla \mathcal{E}(\tilde{X}_t)) dB_t^3) \end{aligned}$$

After taking the expectation, applying Fubini’s theorem and observing that the stochastic integrals of the form $\mathbb{E} \int_0^t \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot D(\cdot) dB_t$ vanish as a consequence of [50, Theorem 3.2.1(iii)] due to the established regularity $\tilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ and $\phi \in C_*^2(\mathbb{R}^d \times \mathbb{R}^d)$, we obtain

$$\begin{aligned} \frac{d}{dt} \mathbb{E} \phi(\tilde{X}_t, \tilde{Y}_t) &= -\mathbb{E} \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (\lambda_1(\tilde{X}_t - u_t) + \lambda_2(\tilde{X}_t - \tilde{Y}_t) + \lambda_3 \nabla \mathcal{E}(\tilde{X}_t)) \\ &\quad + \mathbb{E} \nabla_y \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (\kappa(\tilde{X}_t - \tilde{Y}_t) S^{\beta, \theta}(\tilde{X}_t, \tilde{Y}_t)) \\ &\quad + \frac{1}{2} \sum_{k=1}^d \partial_{x_k x_k}^2 \phi(\tilde{X}_t, \tilde{Y}_t) \left(\sigma_1^2 D(\tilde{X}_t - u_t)_{kk}^2 + \sigma_2^2 D(\tilde{X}_t - \tilde{Y}_t)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(\tilde{X}_t))_{kk}^2 \right) \end{aligned}$$

according to the fundamental theorem of calculus. This shows that $\tilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ satisfies the Fokker-Planck equation

$$\begin{aligned} \frac{d}{dt} \iint \phi(x, y) d\tilde{\rho}_t(x, y) &= - \iint \kappa S^{\beta, \theta}(x, y) \langle y - x, \nabla_y \phi(x, y) \rangle d\tilde{\rho}_t(x, y) \\ &\quad - \iint \lambda_1 \langle x - u_t, \nabla_x \phi(x, y) \rangle + \lambda_2 \langle x - y, \nabla_x \phi(x, y) \rangle + \lambda_3 \langle \nabla \mathcal{E}(x), \nabla_x \phi(x, y) \rangle d\tilde{\rho}_t(x, y) \quad (2.7) \\ &\quad + \frac{1}{2} \iint \sum_{k=1}^d (\sigma_1^2 D(x - u_t)_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(x))_{kk}^2) \partial_{x_k x_k}^2 \phi(x, y) d\tilde{\rho}_t(x, y) \end{aligned}$$

The remainder is identical to the cited reference and is summarised below for completeness.

Step 3: Setting $\mathcal{T}u := y_\alpha(\tilde{\rho}_Y) \in \mathcal{C}([0, T], \mathbb{R}^d)$ provides the self-mapping property of the map

$$\mathcal{T} : \mathcal{C}([0, T], \mathbb{R}^d) \rightarrow \mathcal{C}([0, T], \mathbb{R}^d), \quad u \mapsto \mathcal{T}u = y_\alpha(\tilde{\rho}_Y),$$

which is compact as a consequence of a stability estimate for the consensus point [12, Lemma 3.2]. More precisely, as shown in the cited result, it holds $\|y_\alpha(\tilde{\rho}_{Y,t}) - y_\alpha(\tilde{\rho}_{Y,s})\|_2 \lesssim W_2(\tilde{\rho}_{Y,t}, \tilde{\rho}_{Y,s})$ for $\tilde{\rho}_{Y,t}, \tilde{\rho}_{Y,s} \in \mathcal{P}_4(\mathbb{R}^d)$. Together with the Hölder-1/2 continuity of the Wasserstein-2 distance $W_2(\tilde{\rho}_{Y,t}, \tilde{\rho}_{Y,s})$, this ensures the claimed compactness of \mathcal{T} .

Step 4: Then, for $u = \vartheta \mathcal{T}u$ with $\vartheta \in [0, 1]$, there exists $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ satisfying (2.7) with marginal ρ_Y such that $u = \vartheta y_\alpha(\rho_Y)$. For such u , a uniform bound can be obtained either thanks to the boundedness or the growth condition of \mathcal{E} required in the statement. An application of the Leray-Schauder fixed point theorem concludes the proof by providing a solution to (1.5). \square

2.2. Main result

We now present the main theoretical result about global mean-field law convergence of CBO with memory effects and gradient information for objectives that satisfy the following conditions.

Definition 2.4 (Assumptions). Throughout, we are interested in functions $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$, for which

A1 there exists a unique $x^* \in \mathbb{R}^d$ such that $\mathcal{E}(x^*) = \inf_{x \in \mathbb{R}^d} \mathcal{E}(x) =: \underline{\mathcal{E}}$, and

A2 there exist $\mathcal{E}_\infty, R_0, \eta > 0$, and $v \in (0, \infty)$ such that

$$\|x - x^*\|_\infty \leq \frac{1}{\eta} (\mathcal{E}(x) - \underline{\mathcal{E}})^v \quad \text{for all } x \in B_{R_0}^\infty(x^*), \quad (2.8)$$

$$\mathcal{E}_\infty < \mathcal{E}(x) - \underline{\mathcal{E}} \quad \text{for all } x \in (B_{R_0}^\infty(x^*))^c. \quad (2.9)$$

Furthermore, for the case of an additional gradient drift component, i.e., if $\lambda_3 \neq 0$, we additionally require that $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ and that

A3 there exist $C_{\nabla\mathcal{E}} > 0$ such that

$$\|\nabla\mathcal{E}(x)\|_2 \leq C_{\nabla\mathcal{E}} \|x - x^*\|_2 \quad \text{for all } x \in \mathbb{R}^d. \tag{2.10}$$

In the case, where no gradient drift is present, i.e., $\lambda_3 = 0$ in equations (1.1), (1.5) and (1.6), the objective function \mathcal{E} is only required to be continuous and satisfy Assumptions A1 and A2. While the former merely imposes that the infimum is attained at x^* , the latter can be regarded as a tractability condition of the energy landscape of \mathcal{E} [26, 28]. More precisely, the inverse continuity condition (2.8) ensures that \mathcal{E} is locally coercive in some neighbourhood of the global minimiser x^* . Condition (2.9), on the other hand, guarantees that in the farfield \mathcal{E} is bounded away from the minimal value by at least \mathcal{E}_∞ . This in particular excludes objectives for which $\mathcal{E}(x) \approx \mathcal{E}$ far away from x^* . Note that A2 actually already implies the uniqueness of x^* requested in A1. In case of an additional gradient drift term in the dynamics, i.e., $\lambda_3 \neq 0$, the objective naturally needs to be continuously differentiable. Furthermore, in Assumption A3 we impose that the gradient $\nabla\mathcal{E}$ grows at most linearly. This is a significantly weaker assumption compared to typical smoothness assumptions about \mathcal{E} in the optimisation literature (in particular in the analysis of stochastic gradient descent), where Lipschitz-continuity of the gradient of \mathcal{E} is required [49].

We are now ready to state the main theoretical result. Its proof is deferred to Section 3. For the reader's convenience let us recall that

$$W_2^2(\rho_t, \delta_{(x^*, x^*)}) = \iint (\|x - x^*\|_2^2 + \|y - x^*\|_2^2) d\rho_t(x, y),$$

which motivates to investigate the behaviour of the Lyapunov functional $\mathcal{V}(\rho_t)$ as introduced in (2.11) below.

Theorem 2.5. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 and A2. Furthermore, in the case of an active gradient drift in the CBO dynamics (1.5), i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ obey in addition A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$ be such that $(x^*, x^*) \in \text{supp}(\rho_0)$. Let us define the functional*

$$\mathcal{V}(\rho_t) := \frac{1}{2} \iint (\|x - x^*\|_2^2 + \|y - x\|_2^2) d\rho_t(x, y), \tag{2.11}$$

and the rates

$$\chi_1 := \min \{ \lambda_1 - \lambda_2 - 3\lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_1^2 - 2\sigma_3^2 C_{\nabla\mathcal{E}}^2, 2\kappa\theta + \lambda_2 - \lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_2^2 \}, \tag{2.12a}$$

$$\chi_2 := \max \{ 3\lambda_1 + \lambda_2 + 3\lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_1^2 + 2\sigma_3^2 C_{\nabla\mathcal{E}}^2, 2\kappa\theta + 3\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_2^2 \}, \tag{2.12b}$$

which we assume to be strictly positive through a sufficient choice of the parameters of the CBO dynamics. Furthermore, provided that $\mathcal{V}(\rho_0) > 0$, fix any $\varepsilon \in (0, \mathcal{V}(\rho_0))$, $\vartheta \in (0, 1)$ and define the time horizon

$$T^* := \frac{1}{(1 - \vartheta)\chi_1} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right). \tag{2.13}$$

Then, there exists $\alpha_0 > 0$, depending (among problem-dependent quantities) also on ε and ϑ , such that for all $\alpha > \alpha_0$, if $\rho \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (1.6) on the time interval $[0, T^*]$ with initial condition ρ_0 , we have

$$\mathcal{V}(\rho_T) = \varepsilon \quad \text{with} \quad T \in \left[\frac{(1 - \vartheta)\chi_1}{(1 + \vartheta/2)\chi_2} T^*, T^* \right]. \tag{2.14}$$

Furthermore, on the time interval $[0, T]$, $\mathcal{V}(\rho_t)$ decays at least exponentially fast, i.e., for all $t \in [0, T]$ it holds

$$W_2^2(\rho_t, \delta_{(x^*, x^*)}) \leq 6\mathcal{V}(\rho_t) \leq 6\mathcal{V}(\rho_0) \exp(-(1 - \vartheta)\chi_1 t). \tag{2.15}$$

Theorem 2.5 proves the exponentially fast convergence of the law ρ of the dynamics (1.5) to the global minimiser x^* of \mathcal{E} under a minimal assumption about the initial distribution ρ_0 . The result in particular allows to devise a strategy for the parameter choices of the method. Namely, fixing the parameters $\lambda_2, \lambda_3, \sigma_1, \sigma_2, \sigma_3$ and θ , choosing λ_1 and consecutively κ such that

$$\lambda_1 > \lambda_2 + 3\lambda_3 C_{\nabla \mathcal{E}} + 2\sigma_1^2 + 2\sigma_3^2 C_{\nabla \mathcal{E}}^2 \quad \text{and} \quad \kappa > \frac{1}{2\theta} (-\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla \mathcal{E}} + 2\sigma_2^2)$$

ensures that the convergence rate χ_1 is strictly positive. Since $\chi_2 \geq \chi_1, \chi_2 > 0$ as well. Given a desired accuracy ε , by consulting the proof in Section 3.4, we can further derive an estimate on the lower bound of α , namely

$$\alpha_0 \sim d + \log 16d + \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) - \log c(\vartheta, \chi_1, \lambda_1, \sigma_1) - \log \rho_0(B_r^\infty(x^*) \times B_r^\infty(x^*))$$

for some suitably small $r \in (0, R_0)$, which, like the hidden constant, may depend on ε . The choice of the first set of parameters, in particular what concerns the drift towards the historical best and in the direction of the negative gradient, requires some manual hyperparameter tuning and depends on the problem at hand. We will see this also in Section 4, where we conduct numerical experiments in different application areas.

Eventually, with (2.13) one can determine the maximal time horizon T^* , until which the Lyapunov functional $\mathcal{V}(\rho_t)$ is guaranteed to have reached the prescribed ε . The exact time point T , where $\mathcal{V}(\rho_T) = \varepsilon$, is characterised more concretely in equation (2.14). Due to the presence of memory effects and gradient information, which might counteract the consensus drift of CBO, it seems challenging to specify T more closely. However, in the case of standard CBO, T turns out to be equal to T^* up to a factor depending merely on ϑ , see, e.g., [28].

In fact, this result can be retrieved as a special case of the subsequent Corollary 2.6, where we state an analogous convergence result for the CBO dynamics with gradient information but without memory effect. Its respective proof follows the lines of the one of the richer dynamics in Section 3, cf. also [28, Theorem 12] and [29, Theorem 2], and it is left as an exercise to the reader interested in the technical details of the proof technique. More precisely, for the instantaneous CBO model with gradient drift,

$$d\tilde{X}_t^i = -\lambda_1(\tilde{X}_t^i - y_\alpha(\tilde{\rho}_t^N)) dt - \lambda_3 \nabla \mathcal{E}(\tilde{X}_t^i) dt + \sigma_1 D(\tilde{X}_t^i - y_\alpha(\tilde{\rho}_t^N)) d\tilde{B}_t^{1,i} + \sigma_3 D(\nabla \mathcal{E}(\tilde{X}_t^i)) d\tilde{B}_t^{3,i}, \tag{2.16}$$

where $\tilde{\rho}_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{X}_t^i}$ and to which the associated mean-field Fokker-Planck equation reads

$$\partial_t \tilde{\rho}_t = \operatorname{div}((\lambda_1(x - y_\alpha(\tilde{\rho}_t)) + \lambda_3 \nabla \mathcal{E}(x)) \tilde{\rho}_t) + \frac{1}{2} \sum_{k=1}^d \partial_{x_k x_k}^2 ((\sigma_1^2 D(x - y_\alpha(\tilde{\rho}_t))_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(x))_{kk}^2) \tilde{\rho}_t), \tag{2.17}$$

we have the following convergence result.

Corollary 2.6. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 and A2. Furthermore, in the case of an active gradient drift, i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ obey in addition A3. Moreover, let $\tilde{\rho}_0 \in \mathcal{P}_4(\mathbb{R}^d)$ be such that $x^* \in \operatorname{supp}(\tilde{\rho}_0)$. Let us define the functional*

$$\tilde{\mathcal{V}}(\tilde{\rho}_t) := \frac{1}{2} \int \|x - x^*\|_2^2 d\tilde{\rho}_t(x), \tag{2.18}$$

and the rates

$$\tilde{\chi}_1 := 2\lambda_1 - 2\lambda_3 C_{\nabla \mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla \mathcal{E}}^2, \tag{2.19a}$$

$$\tilde{\chi}_2 := 2\lambda_1 + 2\lambda_3 C_{\nabla \mathcal{E}} - \sigma_1^2 + \sigma_3^2 C_{\nabla \mathcal{E}}^2, \tag{2.19b}$$

which we assume to be strictly positive through a sufficient choice of the parameters of the CBO dynamics. Furthermore, provided that $\tilde{\mathcal{V}}(\tilde{\rho}_0) > 0$, fix any $\varepsilon \in (0, \tilde{\mathcal{V}}(\tilde{\rho}_0))$, $\vartheta \in (0, 1)$ and define the time horizon

$$\tilde{T}^* := \frac{1}{(1 - \vartheta)\tilde{\chi}_1} \log \left(\frac{\tilde{\mathcal{V}}(\tilde{\rho}_0)}{\varepsilon} \right). \tag{2.20}$$

Then, there exists $\tilde{\alpha}_0 > 0$, depending (among problem-dependent quantities) also on ε and ϑ , such that for all $\alpha > \tilde{\alpha}_0$, if $\tilde{\rho} \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (2.17) on the time interval $[0, \tilde{T}^*]$ with initial condition $\tilde{\rho}_0$, we have

$$\tilde{\mathcal{V}}(\tilde{\rho}_{\tilde{T}}) = \varepsilon \quad \text{with} \quad \tilde{T} \in \left[\frac{(1 - \vartheta)\tilde{\chi}_1}{(1 + \vartheta/2)\tilde{\chi}_2} \tilde{T}^*, \tilde{T}^* \right]. \tag{2.21}$$

Furthermore, on the time interval $[0, \tilde{T}]$, $\tilde{\mathcal{V}}(\tilde{\rho}_t)$ decays at least exponentially fast, i.e., for all $t \in [0, \tilde{T}]$ it holds

$$W_2^2(\tilde{\rho}_t, \delta_{x^*}) = 2\tilde{\mathcal{V}}(\tilde{\rho}_t) \leq 2\tilde{\mathcal{V}}(\tilde{\rho}_0) \exp(-(1 - \vartheta)\tilde{\chi}_1 t). \tag{2.22}$$

3. Proof details for Section 2.2

In what follows, we provide the proof details for the global mean-field law convergence result of CBO with memory effects and gradient information, Theorem 2.5. The entire section can be read as a proof sketch with Corollaries 3.3 and 3.5, Propositions 3.6 and 3.8 containing the key individual statements. How to combine these results rigorously to complete the proof of Theorem 2.5 is then covered in detail in Section 3.4.

Remark 3.1. Without loss of generality, we assume $\underline{\mathcal{E}} = 0$ throughout this section.

3.1. Evolution of the mean-field limit

Recall that our overall goal is to establish the convergence of the dynamics (1.6) to a Dirac delta at the global minimiser x^* with respect to the Wasserstein-2 distance, i.e.,

$$W_2(\rho_t, \delta_{(x^*, x^*)}) \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty.$$

To this end, we analyse the decay behaviour of the functional $\mathcal{V}(\rho_t)$ as defined in (2.11), i.e., $\mathcal{V}(\rho_t) = \frac{1}{2} \iint (\|x - x^*\|_2^2 + \|y - x\|_2^2) d\rho_t(x, y)$. More precisely, we will show its exponential decay with a rate controllable through the parameters of the CBO method.

Let us start below with deriving the evolution inequalities for the functionals

$$\mathcal{X}(\rho_t) = \frac{1}{2} \iint \|x - x^*\|_2^2 d\rho_t(x, y) \quad \text{and} \quad \mathcal{Y}(\rho_t) = \frac{1}{2} \iint \|y - x\|_2^2 d\rho_t(x, y).$$

Lemma 3.2. Let $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, and fix $\alpha, \lambda_1, \sigma_1 > 0$ and $\lambda_2, \sigma_2, \lambda_3, \sigma_3, \beta, \kappa, \theta \geq 0$. Moreover, let $T > 0$ and let $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ be a weak solution to the Fokker-Planck equation (1.6). Then, the functionals $\mathcal{X}(\rho_t)$ and $\mathcal{Y}(\rho_t)$ satisfy

$$\begin{aligned} \frac{d}{dt} \begin{pmatrix} \mathcal{X}(\rho_t) \\ \mathcal{Y}(\rho_t) \end{pmatrix} \leq & - \begin{pmatrix} 2\lambda_1 - \lambda_2 - 2\lambda_3 C_{\nabla \mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla \mathcal{E}}^2 & -\lambda_2 - \sigma_2^2 \\ -\lambda_1 - \lambda_3 C_{\nabla \mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla \mathcal{E}}^2 & 2\kappa\theta + 2\lambda_2 - \lambda_1 - \lambda_3 C_{\nabla \mathcal{E}} - \sigma_2^2 \end{pmatrix} \begin{pmatrix} \mathcal{X}(\rho_t) \\ \mathcal{Y}(\rho_t) \end{pmatrix} \\ & + \sqrt{2} \begin{pmatrix} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{X}(\rho_t)} \\ \lambda_1 \sqrt{\mathcal{Y}(\rho_t)} + \sigma_1^2 \sqrt{\mathcal{X}(\rho_t)} \end{pmatrix} \|y_{\alpha}(\rho_{Y,t}) - x^*\|_2 + \frac{\sigma_1^2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \|y_{\alpha}(\rho_{Y,t}) - x^*\|_2^2, \end{aligned}$$

where the inequality has to be understood component-wise.

Proof. We note that the functions $\phi_{\mathcal{X}}(x, y) = 1/2 \|x - x^*\|_2^2$ and $\phi_{\mathcal{Y}}(x, y) = 1/2 \|y - x\|_2^2$ are in $C_*^2(\mathbb{R}^d \times \mathbb{R}^d)$ and recall that ρ satisfies the weak solution identity (2.1) for such test functions. Hence, by applying (2.1) with $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ as above, we obtain for the evolution of $\mathcal{X}(\rho_t)$

$$\begin{aligned} \frac{d}{dt} \mathcal{X}(\rho_t) &= - \iint \lambda_1 \langle x - y_\alpha(\rho_{Y,t}), x - x^* \rangle + \lambda_2 \langle x - y, x - x^* \rangle + \lambda_3 \langle \nabla \mathcal{E}(x), x - x^* \rangle d\rho_t(x, y) \\ &\quad + \frac{1}{2} \iint \sigma_1^2 \|x - y_\alpha(\rho_{Y,t})\|_2^2 + \sigma_2^2 \|x - y\|_2^2 + \sigma_3^2 \|\nabla \mathcal{E}(x)\|_2^2 d\rho_t(x, y) \end{aligned} \tag{3.1}$$

and for the evolution of $\mathcal{Y}(\rho_t)$

$$\begin{aligned} \frac{d}{dt} \mathcal{Y}(\rho_t) &= - \iint \kappa S^{\beta, \theta}(x, y) \|x - y\|_2^2 d\rho_t(x, y) \\ &\quad - \iint \lambda_1 \langle x - y_\alpha(\rho_{Y,t}), x - y \rangle + \lambda_2 \|x - y\|_2^2 + \lambda_3 \langle \nabla \mathcal{E}(x), x - y \rangle d\rho_t(x, y) \\ &\quad + \frac{1}{2} \iint \sigma_1^2 \|x - y_\alpha(\rho_{Y,t})\|_2^2 + \sigma_2^2 \|x - y\|_2^2 + \sigma_3^2 \|\nabla \mathcal{E}(x)\|_2^2 d\rho_t(x, y). \end{aligned} \tag{3.2}$$

Here we used $\nabla_x \phi_{\mathcal{X}}(x, y) = x - x^*$, $\nabla_y \phi_{\mathcal{X}}(x, y) = 0$, $\partial_{x_k x_k}^2 \phi_{\mathcal{X}}(x, y) = 1$, $\nabla_x \phi_{\mathcal{Y}}(x, y) = x - y$, $\nabla_y \phi_{\mathcal{Y}}(x, y) = y - x$ and $\partial_{x_k x_k}^2 \phi_{\mathcal{Y}}(x, y) = 1$. Let us now collect auxiliary estimates in (3.3a)–(3.3g), which turn out to be useful in establishing upper bounds for (3.1) and (3.2). Using standard tools such as Cauchy-Schwarz and Young’s inequality we have

$$- \langle x - y, x - x^* \rangle \leq \|x - y\|_2 \|x - x^*\|_2 \leq \frac{1}{2} (\|x - y\|_2^2 + \|x - x^*\|_2^2), \tag{3.3a}$$

$$\begin{aligned} - \langle x - y_\alpha(\rho_{Y,t}), x - x^* \rangle &= - \|x - x^*\|_2^2 - \langle x^* - y_\alpha(\rho_{Y,t}), x - x^* \rangle \\ &\leq - \|x - x^*\|_2^2 + \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - x^*\|_2, \end{aligned} \tag{3.3b}$$

$$\begin{aligned} - \langle x - y_\alpha(\rho_{Y,t}), x - y \rangle &= - \langle x - x^*, x - y \rangle - \langle x^* - y_\alpha(\rho_{Y,t}), x - y \rangle \\ &\leq \frac{1}{2} (\|x - y\|_2^2 + \|x - x^*\|_2^2) + \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - y\|_2, \end{aligned} \tag{3.3c}$$

$$\begin{aligned} \|x - y_\alpha(\rho_{Y,t})\|_2^2 &= \|x - x^*\|_2^2 - 2 \langle y_\alpha(\rho_{Y,t}) - x^*, x - x^* \rangle + \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2 \\ &\leq \|x - x^*\|_2^2 + 2 \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - x^*\|_2 + \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2, \end{aligned} \tag{3.3d}$$

where in (3.3b)–(3.3d) we expanded the left-hand side of the scalar product and the norm by subtracting and adding x^* . Furthermore, by means of A3 we obtain

$$- \langle \nabla \mathcal{E}(x), x - x^* \rangle \leq \|\nabla \mathcal{E}(x)\|_2 \|x - x^*\|_2 \leq C_{\nabla \mathcal{E}} \|x - x^*\|_2^2, \tag{3.3e}$$

$$\begin{aligned} - \langle \nabla \mathcal{E}(x), x - y \rangle &\leq \|\nabla \mathcal{E}(x)\|_2 \|x - y\|_2 \leq C_{\nabla \mathcal{E}} \|x - x^*\|_2 \|x - y\|_2 \\ &\leq \frac{C_{\nabla \mathcal{E}}}{2} (\|x - y\|_2^2 + \|x - x^*\|_2^2), \end{aligned} \tag{3.3f}$$

$$\|\nabla \mathcal{E}(x)\|_2^2 \leq C_{\nabla \mathcal{E}}^2 \|x - x^*\|_2^2. \tag{3.3g}$$

Integrating the bounds (3.3a), (3.3b), (3.3d), (3.3e) and (3.3g) into equation (3.1) results in the upper bound

$$\begin{aligned} \frac{d}{dt} \mathcal{X}(\rho_t) &\leq - (2\lambda_1 - \lambda_2 - 2\lambda_3 C_{\nabla \mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla \mathcal{E}}^2) \mathcal{X}(\rho_t) + (\lambda_2 + \sigma_2^2) \mathcal{Y}(\rho_t) \\ &\quad + \sqrt{2} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{X}(\rho_t)} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 + \frac{\sigma_1^2}{2} \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2, \end{aligned}$$

where we furthermore used that by Jensen’s inequality

$$\iint \|x - x^*\|_2 d\rho_t(x, y) \leq \sqrt{\iint \|x - x^*\|_2^2 d\rho_t(x, y)} = \sqrt{2\mathcal{X}(\rho_t)}. \tag{3.4}$$

For equation (3.2), we first note that, by definition, $S^{\beta,\theta} \geq \theta$ uniformly. This combined with the bounds (3.3c), (3.3d), (3.3f) and (3.3g) allows to derive

$$\begin{aligned} \frac{d}{dt} \mathcal{Y}(\rho_t) \leq & - (2\kappa\theta + 2\lambda_2 - \lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - \sigma_2^2) \mathcal{Y}(\rho_t) + (\lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} + \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2) \mathcal{X}(\rho_t) \\ & + \sqrt{2} \left(\lambda_1 \sqrt{\mathcal{Y}(\rho_t)} + \sigma_1^2 \sqrt{\mathcal{X}(\rho_t)} \right) \|y_\alpha(\rho_{Y,t}) - x^*\|_2 + \frac{\sigma_1^2}{2} \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2, \end{aligned}$$

where we used (3.4) together with an analogous bound for $\iint \|x - y\|_2 d\rho_t(x, y)$. □

Recalling that $\mathcal{V}(\rho_t) = \mathcal{X}(\rho_t) + \mathcal{Y}(\rho_t)$ immediately allows to obtain an evolution inequality for $\mathcal{V}(\rho_t)$ of the following form.

Corollary 3.3. *Under the assumptions of Lemma 3.2, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -\chi_1 \mathcal{V}(\rho_t) + 2\sqrt{2} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{V}(\rho_t)} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 + \sigma_1^2 \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2, \tag{3.5}$$

with χ_1 as specified in (2.12a).

Analogously to the upper bounds on the time evolutions of the functionals $\mathcal{X}(\rho_t)$, $\mathcal{Y}(\rho_t)$ and $\mathcal{V}(\rho_t)$, we can derive bounds from below as follows.

Lemma 3.4. *Under the assumptions of Lemma 3.2, the functionals $\mathcal{X}(\rho_t)$ and $\mathcal{Y}(\rho_t)$ satisfy*

$$\begin{aligned} \frac{d}{dt} \begin{pmatrix} \mathcal{X}(\rho_t) \\ \mathcal{Y}(\rho_t) \end{pmatrix} \geq & - \begin{pmatrix} 2\lambda_1 + \lambda_2 + 2\lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2 & \lambda_2 - \sigma_2^2 \\ \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2 & 2\kappa\theta + 2\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - \sigma_2^2 \end{pmatrix} \begin{pmatrix} \mathcal{X}(\rho_t) \\ \mathcal{Y}(\rho_t) \end{pmatrix} \\ & - \sqrt{2} \begin{pmatrix} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{X}(\rho_t)} \\ \lambda_1 \sqrt{\mathcal{Y}(\rho_t)} + \sigma_1^2 \sqrt{\mathcal{X}(\rho_t)} \end{pmatrix} \|y_\alpha(\rho_{Y,t}) - x^*\|_2, \end{aligned}$$

where the inequality has to be understood component-wise.

Proof. By following the lines of the proof of Lemma 3.2 and noticing that in analogy to the estimates (3.3), it hold

$$-\langle x - y, x - x^* \rangle \geq -\|x - y\|_2 \|x - x^*\|_2 \geq -\frac{1}{2} (\|x - y\|_2^2 + \|x - x^*\|_2^2), \tag{3.6a}$$

$$\begin{aligned} -\langle x - y_\alpha(\rho_{Y,t}), x - x^* \rangle &= -\|x - x^*\|_2^2 - \langle x^* - y_\alpha(\rho_{Y,t}), x - x^* \rangle \\ &\geq -\|x - x^*\|_2^2 - \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - x^*\|_2, \end{aligned} \tag{3.6b}$$

$$\begin{aligned} -\langle x - y_\alpha(\rho_{Y,t}), x - y \rangle &= -\langle x - x^*, x - y \rangle - \langle x^* - y_\alpha(\rho_{Y,t}), x - y \rangle \\ &\geq -\frac{1}{2} (\|x - y\|_2^2 + \|x - x^*\|_2^2) - \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - y\|_2, \end{aligned} \tag{3.6c}$$

$$\begin{aligned} \|x - y_\alpha(\rho_{Y,t})\|_2^2 &= \|x - x^*\|_2^2 - 2\langle y_\alpha(\rho_{Y,t}) - x^*, x - x^* \rangle + \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2 \\ &\geq \|x - x^*\|_2^2 - 2\|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - x^*\|_2, \end{aligned} \tag{3.6d}$$

as well as

$$-\langle \nabla\mathcal{E}(x), x - x^* \rangle \geq -\|\nabla\mathcal{E}(x)\|_2 \|x - x^*\|_2 \geq -C_{\nabla\mathcal{E}} \|x - x^*\|_2^2, \tag{3.6e}$$

$$\begin{aligned} -\langle \nabla\mathcal{E}(x), x - y \rangle &\geq -\|\nabla\mathcal{E}(x)\|_2 \|x - y\|_2 \geq -C_{\nabla\mathcal{E}} \|x - x^*\|_2 \|x - y\|_2 \\ &\geq -\frac{C_{\nabla\mathcal{E}}}{2} (\|x - y\|_2^2 + \|x - x^*\|_2^2), \end{aligned} \tag{3.6f}$$

$$\|\nabla\mathcal{E}(x)\|_2^2 \geq -C_{\nabla\mathcal{E}}^2 \|x - x^*\|_2^2. \tag{3.6g}$$

we obtain the statement by integrating the bounds into equations (3.1) and (3.2). □

Corollary 3.5. *Under the assumptions of Lemma 3.2, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt} \mathcal{V}(\rho_t) \geq -\chi_2 \mathcal{V}(\rho_t) - 2\sqrt{2} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{V}(\rho_t)} \|y_\alpha(\rho_{Y,t}) - x^*\|_2, \tag{3.7}$$

with χ_2 as specified in (2.12b).

In order to be able to apply Grönwall’s inequality to (3.5) and (3.7) with the aim of obtaining estimates of the form $\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0)e^{-(1-\vartheta)\chi_1 t}$ and $\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0)e^{-(1-\vartheta/2)\chi_2 t}$ for some $\chi_1, \chi_2 > 0$ and a suitable $\vartheta \in (0, 1)$, it remains to control the quantity $\|y_\alpha(\rho_{Y,t}) - x^*\|_2$ through the choice of the parameter α . This is the content of the next section.

3.2. Quantitative Laplace principle

The well-known Laplace principle [21, 48, 52] asserts that for any absolutely continuous probability distribution $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with $x^* \in \text{supp}(\varrho)$ it holds

$$\lim_{\alpha \rightarrow \infty} \left(-\frac{1}{\alpha} \log \|\omega_\alpha\|_{L_1(\varrho)} \right) = \mathcal{E}(x^*) = \underline{\mathcal{E}}, \tag{3.8}$$

which allows to infer that the α -weighted measure $\omega_\alpha / \|\omega_\alpha\|_{L_1(\varrho)} \varrho$ is concentrated in a small region around the minimiser x^* , provided that \mathcal{E} attains its minimum at a single point, which is however guaranteed by the inverse continuity property A2.

The asymptotic nature of the result (3.8), however, does not permit to obtain the required quantitative estimates, which is the reason why the authors of [28] proposed a quantitative nonasymptotic variant of the Laplace principle. In the following proposition, cf. [29, Proposition 1], we state this result for the setting of anisotropic noise considered throughout the paper.

Proposition 3.6 ([29, Proposition 1]). *Let $\underline{\mathcal{E}} = 0$, $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$ we define $\mathcal{E}_r := \sup_{y \in B_r^\infty(x^*)} \mathcal{E}(y)$. Then, under the inverse continuity property A2, for any $r \in (0, R_0]$ and $q > 0$ such that $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, we have*

$$\|y_\alpha(\varrho) - x^*\|_2 \leq \frac{\sqrt{d}(q + \mathcal{E}_r)^v}{\eta} + \frac{\sqrt{d} \exp(-\alpha q)}{\varrho(B_r^\infty(x^*))} \int \|y - x^*\|_2 d\varrho(y).$$

Proof. The proof is a mere reformulation of the one of [29, Proposition 1], which is presented in what follows for the sake of completeness.

For any $a > 0$, Markov’s inequality gives $\|\omega_\alpha\|_{L_1(\varrho)} \geq a\varrho(\{y : \exp(-\alpha\mathcal{E}(y)) \geq a\})$. By choosing $a = \exp(-\alpha\mathcal{E}_r)$ and noting that

$$\varrho(\{y \in \mathbb{R}^d : \exp(-\alpha\mathcal{E}(y)) \geq \exp(-\alpha\mathcal{E}_r)\}) = \varrho(\{y \in \mathbb{R}^d : \mathcal{E}(y) \leq \mathcal{E}_r\}) \geq \varrho(B_r^\infty(x^*)),$$

we get $\|\omega_\alpha\|_{L_1(\varrho)} \geq \exp(-\alpha\mathcal{E}_r)\varrho(B_r^\infty(x^*))$. Now let $\tilde{r} \geq r > 0$. With the definition of the consensus point $y_\alpha(\varrho) = \int y \omega_\alpha(y) / \|\omega_\alpha\|_{L_1(\varrho)} d\varrho(y)$ and Jensen’s inequality, we can decompose

$$\begin{aligned} \|y_\alpha(\varrho) - x^*\|_\infty &\leq \int_{B_{\tilde{r}}^\infty(x^*)} \|y - x^*\|_\infty \frac{\omega_\alpha(y)}{\|\omega_\alpha\|_{L_1(\varrho)}} d\varrho(y) \\ &\quad + \int_{(B_{\tilde{r}}^\infty(x^*))^c} \|y - x^*\|_\infty \frac{\omega_\alpha(y)}{\|\omega_\alpha\|_{L_1(\varrho)}} d\varrho(y). \end{aligned}$$

After noticing that the first term is bounded by \tilde{r} since $\|y - x^*\|_\infty \leq \tilde{r}$ for all $y \in B_{\tilde{r}}^\infty(x^*)$, we can continue the former with

$$\begin{aligned} \|y_\alpha(\varrho) - x^*\|_\infty &\leq \tilde{r} + \frac{1}{\exp(-\alpha\mathcal{E}_r)\varrho(B_r^\infty(x^*))} \int_{(B_{\tilde{r}}^\infty(x^*))^c} \|y - x^*\|_\infty \omega_\alpha(y) d\varrho(y) \\ &\leq \tilde{r} + \frac{\exp(-\alpha \inf_{y \in (B_{\tilde{r}}^\infty(x^*))^c} \mathcal{E}(y))}{\exp(-\alpha\mathcal{E}_r)\varrho(B_r^\infty(x^*))} \int_{(B_{\tilde{r}}^\infty(x^*))^c} \|y - x^*\|_\infty d\varrho(y) \\ &= \tilde{r} + \frac{\exp(-\alpha (\inf_{y \in (B_{\tilde{r}}^\infty(x^*))^c} \mathcal{E}(y) - \mathcal{E}_r))}{\varrho(B_r^\infty(x^*))} \int \|y - x^*\|_\infty d\varrho(y), \tag{3.9} \end{aligned}$$

where for the second term we used $\|\omega_\alpha\|_{L_1(\mathcal{Q})} \geq \exp(-\alpha \mathcal{E}_r) \varrho(B_r^\infty(x^*))$ from above. Let us now choose $\tilde{r} = (q + \mathcal{E}_r)^v / \eta$, which satisfies $\tilde{r} \geq r$, since **A2** with $\underline{\mathcal{E}} = 0$ and $r \leq R_0$ implies

$$\tilde{r} = \frac{(q + \mathcal{E}_r)^v}{\eta} \geq \frac{\mathcal{E}_r^v}{\eta} = \frac{\left(\sup_{y \in B_r^\infty(x^*)} \mathcal{E}(y)\right)^v}{\eta} \geq \sup_{y \in B_r^\infty(x^*)} \|y - x^*\|_\infty = r.$$

Furthermore, due to the assumption $q + \mathcal{E}_r \leq \mathcal{E}_\infty$ in the statement we have $\tilde{r} \leq \mathcal{E}_\infty^v / \eta$, which together with the two cases of **A2** with $\underline{\mathcal{E}} = 0$ allows to bound the infimum in (3.9) as follows

$$\inf_{y \in (B_r^\infty(x^*))^c} \mathcal{E}(y) - \mathcal{E}_r \geq \min \left\{ \mathcal{E}_\infty, (\eta \tilde{r})^{\frac{1}{v}} \right\} - \mathcal{E}_r = (\eta \tilde{r})^{\frac{1}{v}} - \mathcal{E}_r = (q + \mathcal{E}_r) - \mathcal{E}_r = q.$$

Inserting this and the definition of \tilde{r} into (3.9), we get the result as a consequence of the norm equivalence $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \sqrt{d} \|\cdot\|_\infty$. □

To eventually apply Proposition 3.6 in the setting of Corollary 3.3, i.e., to upper bound the distance of the consensus point $y_\alpha(\rho_{Y,t})$ to the global minimiser x^* , it remains to ensure that $\rho_{Y,t}(B_r^\infty(x^*))$ is bounded away from 0 for a finite time horizon. We ensure that this is indeed the case in what follows.

3.3. A lower bound for the probability mass $\rho_{Y,t}(B_r^\infty(x^*))$

In this section, for any small radius $r > 0$, we provide a lower bound on the probability mass of $\rho_{Y,t}(B_r^\infty(x^*))$ by defining a mollifier $\phi_r : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ so that

$$\rho_{Y,t}(B_r^\infty(x^*)) = \rho_t(\mathbb{R}^d \times B_r^\infty(x^*)) = \iint_{\mathbb{R}^d \times B_r^\infty(x^*)} 1 \, d\rho_t(x, y) \geq \iint \phi_r(x, y) \, d\rho_t(x, y)$$

and studying the evolution of the right-hand side.

Lemma 3.7. *For $r > 0$ let $\Omega_r := \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \max\{\|x - x^*\|_\infty, \|x - y\|_\infty\} < r/2\}$ and define the mollifier $\phi_r : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ by*

$$\phi_r(x, y) := \begin{cases} \prod_{k=1}^d \exp\left(1 - \frac{(\frac{r}{2})^2}{(\frac{r}{2})^2 - (x - x^*)_k^2}\right) \exp\left(1 - \frac{(\frac{r}{2})^2}{(\frac{r}{2})^2 - (x - y)_k^2}\right), & \text{if } (x, y) \in \Omega_r, \\ 0, & \text{else.} \end{cases}$$

We have that $\text{Im}(\phi_r) = [0, 1]$, $\text{supp}(\phi_r) = \Omega_r \subset B_{r/2}^\infty(x^*) \times B_r^\infty(x^*) \subset \mathbb{R}^d \times B_r^\infty(x^*)$, $\phi_r \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ and

$$\begin{aligned} \partial_{x_k} \phi_r(x, y) &= -\frac{r^2}{2} \left(\frac{(x - x^*)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2\right)^2} + \frac{(x - y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2} \right) \phi_r(x, y), \\ \partial_{y_k} \phi_r(x, y) &= -\frac{r^2}{2} \frac{(y - x)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2} \phi_r(x, y), \\ \partial_{x_k x_k}^2 \phi_r(x, y) &= \frac{r^2}{2} \left(\left(\frac{2 \left(2 (x - x^*)_k^2 - \left(\frac{r}{2}\right)^2 \right) (x - x^*)_k^2 - \left(\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2 \right)^2}{\left(\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2\right)^4} \right) \right. \\ &\quad \left. + \left(\frac{2 \left(2 (x - y)_k^2 - \left(\frac{r}{2}\right)^2 \right) (x - y)_k^2 - \left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2 \right)^2}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^4} \right) \right) \phi_r(x, y). \end{aligned}$$

Proof. It is straightforward to check the properties of ϕ_r as it is a tensor product of classical well-studied mollifiers. □

To keep the notation as concise as possible in what follows, let us introduce the decomposition

$$\partial_{x_k} \phi_r = \delta_{x_k}^* \phi_r + \delta_{x_k}^Y \phi_r \quad \text{and} \quad \partial_{x_k x_k}^2 \phi_r = \delta_{x_k x_k}^{2,*} \phi_r + \delta_{x_k x_k}^{2,Y} \phi_r, \tag{3.10}$$

where

$$\delta_{x_k}^* \phi_r(x, y) = \frac{-\frac{r^2}{2} (x-x^*)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-x^*)_k^2\right)^2} \phi_r(x, y) \quad \text{and} \quad \delta_{x_k}^Y \phi_r(x, y) = \frac{-\frac{r^2}{2} (x-y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \phi_r(x, y)$$

and analogously for $\delta_{x_k x_k}^{2,*} \phi_r$ and $\delta_{x_k x_k}^{2,Y} \phi_r$.

Proposition 3.8. *Let $T > 0$, $r > 0$, and fix parameters $\alpha, \lambda_1, \sigma_1 > 0$ as well as parameters $\lambda_2, \sigma_2, \lambda_3, \sigma_3, \beta, \kappa, \theta \geq 0$ such that $\sigma_2 > 0$ iff $\lambda_2 \neq 0$ and $\sigma_3 > 0$ iff $\lambda_3 \neq 0$. Moreover, assume the validity of Assumption A3 if $\lambda_3 \neq 0$. Let $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ weakly solve the Fokker-Planck equation (1.6) in the sense of Definition 2.1 with initial condition $\rho_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ and for $t \in [0, T]$. Then, for all $t \in [0, T]$ we have*

$$\rho_{Y,t}(B_r^\infty(x^*)) \geq \left(\iint \phi_r(x, y) d\rho_0(x, y) \right) \exp(-pt) \tag{3.11}$$

with

$$p := d \sum_{i=1}^3 \omega_i \left((1 + \mathbb{1}_{i \in \{1,3\}}) \left(\frac{2\lambda_i C_Y \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_i^2 C_Y^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_i^2}{\tilde{c}\sigma_i^2} \right) + \mathbb{1}_{i=2} \frac{\sigma_2^2 c}{(1-c)^4} \right), \tag{3.12}$$

where, for any $B < \infty$ with $\sup_{t \in [0, T]} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \leq B$, $C_Y = C_Y(r, B, d, C_{\nabla \mathcal{E}})$ is as defined in (3.20). Moreover, $\omega_i = \mathbb{1}_{\lambda_i > 0}$ for $i \in \{1, 2, 3\}$ and $c \in (1/2, 1)$ can be any constant that satisfies $(1-c)^2 \leq (2c-1)c$.

Remark 3.9. *In order to ensure a finite decay rate $p < \infty$ in Proposition 3.8, it is crucial to have non-vanishing diffusions $\sigma_1 > 0$, $\sigma_2 > 0$ if $\lambda_2 \neq 0$ and $\sigma_3 > 0$ if $\lambda_3 \neq 0$. As apparent from the formulation of the statement as well as the proof below, σ_2 or σ_3 may be 0 if the corresponding drift parameter, λ_2 or λ_3 , respectively, vanishes.*

Proof of Proposition 3.8. By the definition of the marginal ρ_Y and the properties of the mollifier ϕ_r defined in Lemma 3.7, we have

$$\rho_{Y,t}(B_r^\infty(x^*)) = \rho_t(\mathbb{R}^d \times B_r^\infty(x^*)) \geq \rho_t(\Omega_r) \geq \iint \phi_r(x, y) d\rho_t(x, y).$$

Our strategy is to derive a lower bound for the right-hand side of this inequality. Using the weak solution property of ρ as in Definition 2.1 and the fact that $\phi_r \in \mathcal{C}_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, we obtain

$$\begin{aligned} \frac{d}{dt} \iint \phi_r(x, y) d\rho_t(x, y) &= \sum_{k=1}^d \iint T_k^s(x, y) d\rho_t(x, y) \\ &+ \sum_{k=1}^d \iint (T_{1k}^c(x, y) + T_{2k}^c(x, y) + T_{1k}^\ell(x, y) + T_{2k}^\ell(x, y) + T_{1k}^g(x, y) + T_{2k}^g(x, y)) d\rho_t(x, y), \end{aligned} \tag{3.13}$$

where $T_k^s(x, y) := -\kappa S^{\beta, \theta}(x, y) (y-x)_k \partial_{y_k} \phi_r(x, y)$ and

$$\begin{aligned} T_{1k}^c(x, y) &:= -\lambda_1 (x - y_\alpha(\rho_{Y,t}))_k \partial_{x_k} \phi_r(x, y), & T_{2k}^c(x, y) &:= \frac{\sigma_1^2}{2} (x - y_\alpha(\rho_{Y,t}))_k^2 \partial_{x_k x_k}^2 \phi_r(x, y), \\ T_{1k}^\ell(x, y) &:= -\lambda_2 (x-y)_k \partial_{x_k} \phi_r(x, y), & T_{2k}^\ell(x, y) &:= \frac{\sigma_2^2}{2} (x-y)_k^2 \partial_{x_k x_k}^2 \phi_r(x, y), \\ T_{1k}^g(x, y) &:= -\lambda_3 \partial_{x_k} \mathcal{E}(x) \partial_{x_k} \phi_r(x, y), & T_{2k}^g(x, y) &:= \frac{\sigma_3^2}{2} (\partial_{x_k} \mathcal{E}(x))^2 \partial_{x_k x_k}^2 \phi_r(x, y) \end{aligned}$$

for $k \in \{1, \dots, d\}$. Since the mollifier ϕ_r and its derivatives vanish outside of Ω_r , we restrict our attention to Ω_r , and aim for showing for all $k \in \{1, \dots, d\}$ that

- $T_k^s(x, y) \geq 0$,
- $T_{1k}^c(x, y) + T_{2k}^c(x, y) \geq -p^c \phi_r(x, y)$,
- $T_{1k}^\ell(x, y) + T_{2k}^\ell(x, y) \geq -p^\ell \phi_r(x, y)$,
- $T_{1k}^g(x, y) + T_{2k}^g(x, y) \geq -p^g \phi_r(x, y)$

pointwise for all $(x, y) \in \Omega_r$ with suitable constants $0 \leq p^\ell, p^c, p^g < \infty$.

Term T_k^s : Using the expression for $\partial_{y_k} \phi_r$ from Lemma 3.7 and the fact that $S^{\beta,\theta} \geq \theta/2 \geq 0$, it is easy to see that

$$T_k^s(x, y) = \frac{r^2 \kappa}{2} S^{\beta,\theta}(x, y) \frac{(y-x)_k^2}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \phi_r(x, y) \geq 0. \tag{3.14}$$

Terms $T_{1k}^c + T_{2k}^c, T_{1k}^\ell + T_{2k}^\ell$ and $T_{1k}^g + T_{2k}^g$: We first note that the third inequality from above holds with $p^\ell = 0$ if $\lambda_2 = \sigma_2 = 0$ and the fourth with $p^g = 0$ if $\lambda_3 = \sigma_3 = 0$.

Therefore, in what follows we assume that $\lambda_2, \sigma_2, \lambda_3, \sigma_3 > 0$. In order to lower bound the three terms from above, we arrange the summands by using the abbreviations introduced in (3.10) as follows. For $T_{1k}^c + T_{2k}^c$, we have

$$T_{1k}^c(x, y) + T_{2k}^c(x, y) = -\lambda_1 (x - y_\alpha(\rho_{Y,t}))_k \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma_1^2}{2} (x - y_\alpha(\rho_{Y,t}))_k^2 \delta_{x_k x_k}^{2,*} \phi_r(x, y) \tag{3.15a}$$

$$- \lambda_1 (x - y_\alpha(\rho_{Y,t}))_k \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_1^2}{2} (x - y_\alpha(\rho_{Y,t}))_k^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y), \tag{3.15b}$$

for $T_{1k}^\ell + T_{2k}^\ell$ we have

$$T_{1k}^\ell(x, y) + T_{2k}^\ell(x, y) = -\lambda_2 (x - y)_k \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma_2^2}{2} (x - y)_k^2 \delta_{x_k x_k}^{2,*} \phi_r(x, y) \tag{3.16a}$$

$$- \lambda_2 (x - y)_k \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_2^2}{2} (x - y)_k^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y) \tag{3.16b}$$

and for $T_{1k}^g + T_{2k}^g$ we have

$$T_{1k}^g(x, y) + T_{2k}^g(x, y) = -\lambda_3 \partial_{x_k} \mathcal{E}(x) \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma_3^2}{2} (\partial_{x_k} \mathcal{E}(x))^2 \delta_{x_k x_k}^{2,*} \phi_r(x, y) \tag{3.17a}$$

$$- \lambda_3 \partial_{x_k} \mathcal{E}(x) \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_3^2}{2} (\partial_{x_k} \mathcal{E}(x))^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y). \tag{3.17b}$$

We now treat each of the two-part sums in (3.15a), (3.15b), (3.16a), (3.16b), (3.17a) and (3.17b) separately by employing a technique similar to the one used in the proof of [29, Proposition 2], which was developed originally to prove [28, Proposition 20].

Terms (3.15a), (3.16a) and (3.17a): Owing to their similar structure (in particular with respect to the denominator of the derivatives $\delta_{x_k}^* \phi_r$ and $\delta_{x_k x_k}^{2,*} \phi_r$), we can treat the three sums (3.15a), (3.16a) and (3.17a) simultaneously. Therefore, we consider the general formulation

$$-\lambda \Upsilon_k(x, y) \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \delta_{x_k x_k}^{2,*} \phi_r(x, y) =: T_{1k}^*(x, y) + T_{2k}^*(x, y), \tag{3.18}$$

which matches (3.15a) when $\Upsilon_k(x, y) = (x - y_\alpha(\rho_{Y,t}))_k$, $\lambda = \lambda_1$ and $\sigma = \sigma_1$, (3.16a) when $\Upsilon_k(x, y) = (x - y)_k$, $\lambda = \lambda_2$ and $\sigma = \sigma_2$, and (3.17a) when $\Upsilon_k(x, y) = \partial_{x_k} \mathcal{E}(x)$, $\lambda = \lambda_3$ and $\sigma = \sigma_3$.

To achieve the desired lower bound over Ω_r , we introduce the subsets

$$K_{1k}^* := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : |(x - x^*)_k| > \frac{\sqrt{c}}{2} r \right\}$$

and

$$K_{2k}^* := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : -\lambda \Upsilon_k(x, y) (x - x^*)_k \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \right. \\ \left. > \tilde{c} \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - x^*)_k^2 \right\},$$

where $\tilde{c} := 2c - 1 \in (0, 1)$. For fixed k , we now decompose Ω_r according to

$$\Omega_r = ((K_{1k}^*)^c \cap \Omega_r) \cup (K_{1k}^* \cap (K_{2k}^*)^c \cap \Omega_r) \cup (K_{1k}^* \cap K_{2k}^* \cap \Omega_r).$$

In the following, we treat each of these three subsets separately.

Subset $(K_{1k}^)^c \cap \Omega_r$:* We have $|(x - x^*)_k| \leq \frac{\sqrt{c}}{2}r$ for each $(x, y) \in (K_{1k}^*)^c$, which can be used to independently derive lower bounds for both summands in (3.18). For the first, we insert the expression for $\delta_{x_k}^* \phi_r(x, y)$ to get

$$T_{1k}^*(x, y) = \frac{r^2}{2} \lambda \Upsilon_k(x, y) \frac{(x - x^*)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \phi_r(x, y) \\ \geq -\frac{r^2}{2} \lambda \frac{|\Upsilon_k(x, y)| |(x - x^*)_k|}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \phi_r(x, y) \geq -\frac{2\lambda C_\Upsilon \sqrt{c}}{(1 - c)^2 \frac{r}{2}} \phi_r(x, y) \\ =: -p_1^{*,\Upsilon} \phi_r(x, y), \tag{3.19}$$

where, in the last inequality, we used that $(x, y) \in \Omega_r$, the definition of B and Assumption A3 to get the bound

$$|\Upsilon_k(x, y)| = \begin{cases} |(x - y_\alpha(\rho_{Y,t}))_k| \leq \frac{r}{2} + B, & \text{if } \Upsilon_k(x, y) = (x - y_\alpha(\rho_{Y,t}))_k, \\ |(x - y)_k| \leq \frac{r}{2}, & \text{if } \Upsilon_k(x, y) = (x - y)_k, \\ \begin{cases} |\partial_{x_k} \mathcal{E}(x)| \leq \|\nabla \mathcal{E}(x)\|_2 \leq C_{\nabla \mathcal{E}} \|x - x^*\|_2 \\ \leq C_{\nabla \mathcal{E}} d \|x - x^*\|_\infty \leq C_{\nabla \mathcal{E}} d \frac{r}{2} \end{cases} & \text{if } \Upsilon_k(x, y) = \partial_{x_k} \mathcal{E}(x). \end{cases} \\ \leq \max \left\{ \frac{r}{2} + B, C_{\nabla \mathcal{E}} d \frac{r}{2} \right\} =: C_\Upsilon(r, B, d, C_{\nabla \mathcal{E}}). \tag{3.20}$$

For the second summand, we insert the expression for $\delta_{x_k x_k}^{2,*} \phi_r(x, y)$ to obtain

$$T_{2k}^*(x, y) = \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \delta_{x_k x_k}^{2,*} \phi_r(x, y) \\ = \sigma^2 \left(\frac{r}{2} \right)^2 \Upsilon_k^2(x, y) \frac{2 \left(2(x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2 - \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^4} \phi_r(x, y) \\ \geq -\frac{\sigma^2 C_\Upsilon^2}{(1 - c)^4 \left(\frac{r}{2} \right)^2} \phi_r(x, y) =: -p_2^{*,\Upsilon} \phi_r(x, y), \tag{3.21}$$

where the last inequality uses $\Upsilon_k^2(x, y) \leq C_\Upsilon^2$.

Subset $K_{1k}^ \cap (K_{2k}^*)^c \cap \Omega_r$:* As $(x, y) \in K_{1k}^*$, we have $|(x - x^*)_k| > \frac{\sqrt{c}}{2}r$. We observe that the sum in (3.18) is nonnegative for all (x, y) in this subset whenever

$$\left(-\lambda \Upsilon_k(x, y) (x - x^*)_k + \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \right) \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \\ \leq \sigma^2 \Upsilon_k^2(x, y) \left(2(x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2. \tag{3.22}$$

The first term on the left-hand side in (3.22) can be bounded from above by exploiting that $v \in (K_{2k}^*)^c$ and by using the relation $\tilde{c} = 2c - 1$. More precisely, we have

$$\begin{aligned} & -\lambda \Upsilon_k(x, y) (x - x^*)_k \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \leq \tilde{c} \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - x^*)_k^2 \\ & = (2c - 1) \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - x^*)_k^2 \leq \left(2(x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - x^*)_k^2, \end{aligned}$$

where the last inequality follows since $v \in K_{1k}^*$. For the second term on the left-hand side in (3.22), we can use $(1 - c)^2 \leq (2c - 1)c$ as per assumption, to get

$$\begin{aligned} & \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (1 - c)^2 \left(\frac{r}{2} \right)^4 \\ & \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (2c - 1) \left(\frac{r}{2} \right)^2 c \left(\frac{r}{2} \right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(2(x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2. \end{aligned}$$

Hence, (3.22) holds and we have that (3.18) is uniformly nonnegative on this subset.

Subset $K_{1k}^ \cap K_{2k}^* \cap \Omega_r$:* As $(x, y) \in K_{1k}^*$, we have $|(x - x^*)_k| > \frac{\sqrt{c}}{2} r$. To start with we note that the first summand of (3.18) vanishes whenever $\sigma^2 \Upsilon_k^2(x, y) = 0$, provided $\sigma > 0$, so nothing needs to be done if $\Upsilon_k(x, y) = 0$. Otherwise, if $\sigma^2 \Upsilon_k^2(x, y) > 0$, we exploit $(x, y) \in K_{2k}^*$ to get

$$\begin{aligned} & \frac{\Upsilon_k(x, y) (x - x^*)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \geq \frac{-|\Upsilon_k(x, y)| |(x - x^*)_k|}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \\ & > \frac{2\lambda \Upsilon_k(x, y) (x - x^*)_k}{\tilde{c} \left(\frac{r}{2} \right)^2 \sigma^2 |\Upsilon_k(x, y)| |(x - x^*)_k|} \geq -\frac{8\lambda}{\tilde{c} r^2 \sigma^2}. \end{aligned}$$

Using this, the first summand of (3.18) can be bounded from below by

$$T_{1k}^*(x, y) = \lambda \frac{r^2}{2} \frac{\Upsilon_k(x, y) (x - x^*)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \phi_r(x, y) \geq -\frac{4\lambda^2}{\tilde{c}\sigma^2} \phi_r(x, y) =: -p_3^{*,\Upsilon} \phi_r(x, y). \tag{3.23}$$

For the second summand, the nonnegativity of $\sigma^2 \Upsilon_k^2(x, y)$ implies the nonnegativity, whenever

$$2 \left(2(x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2 \geq \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2.$$

This holds for $v \in K_{1k}^*$, if $2(2c - 1)c \geq (1 - c)^2$ as implied by the assumption.

Term (3.16b): Recall that this term has the structure

$$-\lambda_2 (x - y)_k \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma^2}{2} (x - y)_k^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y) =: T_{1k}^{Y,1}(x, y) + T_{2k}^{Y,1}(x, y). \tag{3.24}$$

We first note that the first summand of (3.24) is always nonnegative since

$$T_{1k}^{Y,1}(x, y) = \lambda_2 \frac{r^2}{2} \frac{(x - y)_k^2}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2} \phi_r(x, y) \geq 0. \tag{3.25}$$

For the second summand of (3.24), a direct computation shows

$$T_{2k}^{Y,1}(x, y) = \sigma^2 \left(\frac{r}{2} \right)^2 (x - y)_k^2 \frac{3(x - y)_k^4 - \left(\frac{r}{2} \right)^4}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^4} \phi_r(x, y),$$

which is nonnegative on the set

$$K_k^Y := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : |(x - y)_k| > \frac{\sqrt{c}}{2} r \right\}$$

for any $c \geq 1/\sqrt{3}$, as ensured by $(1 - c)^2 \leq (2c - 1)c$. On the complement $(K_k^Y)^c$, we have $|(x - y)_k| \leq \frac{\sqrt{c}}{2}r$, which can be used to bound

$$\begin{aligned}
 T_{2k}^{Y,1}(x, y) &= \sigma_2^2 \left(\frac{r}{2}\right)^2 (x - y)_k^2 \frac{3(x - y)_k^4 - \left(\frac{r}{2}\right)^4}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^4} \phi_r(x, y) \\
 &\geq -\frac{\sigma_2^2 c}{(1 - c)^4} \phi_r(x, y) =: -p^{Y, \Upsilon_c} \phi_r(x, y).
 \end{aligned}
 \tag{3.26}$$

Terms (3.15b) and (3.17b): The final two terms to be controlled have again a similar structure of the form

$$-\lambda \Upsilon_k(x, y) \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \delta_{x_k x_k}^{2,Y} \phi_r(x, y) =: T_{1k}^{Y,2}(x, y) + T_{2k}^{Y,2}(x, y),
 \tag{3.27}$$

where we recycle the notation introduced after (3.18), i.e., $\Upsilon_k(x, y) = (x - y_\alpha(\rho_{Y,t}))_k$, $\lambda = \lambda_1$ and $\sigma = \sigma_1$ in the case of (3.15b) and $\Upsilon_k(x, y) = \partial_{x_k} \mathcal{E}(x)$, $\lambda = \lambda_3$ and $\sigma = \sigma_3$ in the case of (3.17b).

The procedure for deriving lower bounds is similar to the one at the beginning with the exception that the denominator of the derivatives $\delta_{x_k}^Y \phi_r$ and $\delta_{x_k x_k}^{2,Y} \phi_r$ requires to introduce an adapted decomposition of Ω_r . To be more specific, we define the subsets

$$K_{1k}^Y := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : |(x - y)_k| > \frac{\sqrt{c}}{2}r \right\}$$

and

$$\begin{aligned}
 K_{2k}^Y := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : -\lambda \Upsilon_k(x, y) (x - y)_k \left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2 \right) \right. \\
 \left. > \tilde{c} \left(\frac{r}{2}\right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - y)_k^2 \right\},
 \end{aligned}$$

where $\tilde{c} := 2c - 1 \in (0, 1)$. For fixed k , we now decompose Ω_r according to

$$\Omega_r = ((K_{1k}^Y)^c \cap \Omega_r) \cup (K_{1k}^Y \cap (K_{2k}^Y)^c \cap \Omega_r) \cup (K_{1k}^Y \cap K_{2k}^Y \cap \Omega_r).$$

In the following, we treat again each of these three subsets separately.

Subset $(K_{1k}^Y)^c \cap \Omega_r$: We have $|(x - y)_k| \leq \frac{\sqrt{c}}{2}r$ for each $(x, y) \in (K_{1k}^Y)^c$, which can be used to independently derive lower bounds for both summands in (3.27). For the first summand, we insert the expression for $\delta_{x_k}^Y \phi_r(x, y)$ to get

$$\begin{aligned}
 T_{1k}^{Y,2}(x, y) &= \frac{r^2}{2} \lambda \Upsilon_k(x, y) \frac{(x - y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2} \phi_r(x, y) \\
 &\geq -\frac{r^2}{2} \lambda \frac{|\Upsilon_k(x, y)| |(x - y)_k|}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2} \phi_r(x, y) \geq -\frac{2\lambda C_\Upsilon \sqrt{c}}{(1 - c)^2 \frac{r}{2}} \phi_r(x, y) \\
 &=: -p_1^{Y, \Upsilon} \phi_r(x, y),
 \end{aligned}
 \tag{3.28}$$

where we recall from above that $\Upsilon_k(x, y) \leq C_\Upsilon$, which was used in the last inequality. For the second summand, we insert the expression for $\delta_{x_k x_k}^{2,Y} \phi_r(x, y)$ to obtain

$$\begin{aligned}
 T_{2k}^{Y,2}(x, y) &= \sigma^2 \left(\frac{r}{2}\right)^2 \Upsilon_k^2(x, y) \frac{2 \left(2(x - y)_k^2 - \left(\frac{r}{2}\right)^2 \right) (x - y)_k^2 - \left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^4} \phi_r(x, y) \\
 &\geq -\frac{\sigma^2 C_\Upsilon^2}{(1 - c)^4 \left(\frac{r}{2}\right)^2} \phi_r(x, y) =: -p_2^{Y, \Upsilon} \phi_r(x, y),
 \end{aligned}
 \tag{3.29}$$

where the last inequality uses $\Upsilon_k^2(x, y) \leq C_\Upsilon^2$.

Subset $K_{1k}^Y \cap (K_{2k}^Y)^c \cap \Omega_r$: As $(x, y) \in K_{1k}^Y$, we have $|(x - y)_k| > \frac{\sqrt{c}}{2}r$. We observe that the sum in (3.27) is nonnegative for all (x, y) in this subset whenever

$$\begin{aligned} & \left(-\lambda \Upsilon_k(x, y) (x - y)_k + \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \right) \left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2 \\ & \leq \sigma^2 \Upsilon_k^2(x, y) \left(2(x - y)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - y)_k^2. \end{aligned} \tag{3.30}$$

The first term on the left-hand side in (3.30) can be bounded from above exploiting that $v \in (K_{2k}^Y)^c$ and by using the relation $\tilde{c} = 2c - 1$. More precisely, we have

$$\begin{aligned} -\lambda \Upsilon_k(x, y) (x - y)_k \left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2 & \leq \tilde{c} \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - y)_k^2 \\ & = (2c - 1) \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - y)_k^2 \leq \left(2(x - y)_k^2 - \left(\frac{r}{2} \right)^2 \right) \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - y)_k^2, \end{aligned}$$

where the last inequality follows since $v \in K_{1k}^Y$. For the second term on the left-hand side in (3.30), we can use $(1 - c)^2 \leq (2c - 1)c$ as per assumption, to get

$$\begin{aligned} \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2 & \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (1 - c)^2 \left(\frac{r}{2} \right)^4 \\ & \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (2c - 1) \left(\frac{r}{2} \right)^2 c \left(\frac{r}{2} \right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(2(x - y)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - y)_k^2. \end{aligned}$$

Hence, (3.30) holds and we have that (3.27) is uniformly nonnegative on this subset.

Subset $K_{1k}^Y \cap K_{2k}^Y \cap \Omega_r$: As $(x, y) \in K_{1k}^Y$, we have $|(x - y)_k| > \frac{\sqrt{c}}{2}r$. To start with we note that the first summand of (3.27) vanishes whenever $\sigma^2 \Upsilon_k^2(x, y) = 0$, provided $\sigma > 0$, so nothing needs to be done if $\Upsilon_k(x, y) = 0$. Otherwise, if $\sigma^2 \Upsilon_k^2(x, y) > 0$, we exploit $(x, y) \in K_{2k}^Y$ to get

$$\begin{aligned} \frac{\Upsilon_k(x, y) (x - y)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2} & \geq \frac{-|\Upsilon_k(x, y)| |(x - y)_k|}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2} \\ & > \frac{2\lambda \Upsilon_k(x, y) (x - y)_k}{\tilde{c} \left(\frac{r}{2} \right)^2 \sigma^2 |\Upsilon_k(x, y)| |(x - y)_k|} \geq -\frac{8\lambda}{\tilde{c} r^2 \sigma^2}. \end{aligned}$$

Using this, the first summand of (3.27) can be bounded from below by

$$T_{1k}^{Y,2}(x, y) = \lambda \frac{r^2}{2} \frac{\Upsilon_k(x, y) (x - y)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2} \phi_r(x, y) \geq -\frac{4\lambda^2}{\tilde{c}\sigma^2} \phi_r(x, y) =: -p_3^{Y,r} \phi_r(x, y). \tag{3.31}$$

For the second summand, the nonnegativity of $\sigma^2 \Upsilon_k^2(x, y)$ implies the nonnegativity, whenever

$$2 \left(2(x - y)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - y)_k^2 \geq \left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2.$$

This holds for $v \in K_{1k}^Y$, if $2(2c - 1)c \geq (1 - c)^2$ as implied by the assumption.

Concluding the proof: Combining the formerly established lower bounds (3.19), (3.21), (3.23), (3.25), (3.26), (3.28), (3.29) and (3.31), we obtain for the constants p^c, p^ℓ and p^s defined at the beginning of the proof

$$\begin{aligned}
 p^c &= p_1^{*,\Upsilon_c} + p_2^{*,\Upsilon_c} + p_3^{*,\Upsilon_c} + p_1^{Y,\Upsilon_c} + p_2^{Y,\Upsilon_c} + p_3^{Y,\Upsilon_c} = 2 \left(\frac{2\lambda_1 C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_1^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_1^2}{\tilde{c}\sigma_1^2} \right) \\
 p^\ell &= p_1^{*,\Upsilon_\ell} + p_2^{*,\Upsilon_\ell} + p_3^{*,\Upsilon_\ell} + p_1^{Y,\Upsilon_\ell} = \frac{2\lambda_2 C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_2^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_2^2}{\tilde{c}\sigma_2^2} + \frac{\sigma_2^2 c}{(1-c)^4} \\
 p^g &= p_1^{*,\Upsilon_g} + p_2^{*,\Upsilon_g} + p_3^{*,\Upsilon_g} + p_1^{Y,\Upsilon_g} + p_2^{Y,\Upsilon_g} + p_3^{Y,\Upsilon_g} = 2 \left(\frac{2\lambda_3 C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_3^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_3^2}{\tilde{c}\sigma_3^2} \right).
 \end{aligned}
 \tag{3.32}$$

Together with (3.14) and by using the evolution of ϕ_r as in (3.13), we eventually obtain

$$\begin{aligned}
 \frac{d}{dt} \iint \phi_r d\rho_t &\geq -d(p^c + p^\ell + p^g) \iint \phi_r d\rho_t \\
 &\geq -d \sum_{i=1}^3 \omega_i \left((1 + \mathbb{1}_{i \neq 2}) \left(\frac{2\lambda_i C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_i^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_i^2}{\tilde{c}\sigma_i^2} \right) + \mathbb{1}_{i=2} \frac{\sigma_2^2 c}{(1-c)^4} \right) \iint \phi_r d\rho_t \\
 &= -q \iint \phi_r d\rho_t,
 \end{aligned}$$

where q is defined implicitly and where $\omega_i = \mathbb{1}_{\lambda_i > 0}$ for $i \in \{1, 2, 3\}$. Notice that $\omega_1 = 1$ since $\lambda_1 > 0$ by assumption. An application of Grönwall’s inequality concludes the proof. \square

3.4. Proof of Theorem 2.5

We now have all necessary tools at hand to prove the global mean-field law convergence result for CBO with memory effects and gradient information by rigorously combining the formerly discussed statements.

Proof of Theorem 2.5. If $\mathcal{V}(\rho_0) = 0$, there is nothing to be shown since in this case $\rho_0 = \delta_{(x^*, x^*)}$. Thus, let $\mathcal{V}(\rho_0) > 0$ in what follows.

Let us first choose the parameter α such that

$$\alpha > \alpha_0 := \frac{1}{q_\varepsilon} \left(\log \left(\frac{2^{d+2} \sqrt{d}}{c(\vartheta, \chi_1, \lambda_1, \sigma_1)} \right) + \max \left\{ \frac{1}{2}, \frac{p}{(1-\vartheta)\chi_1} \right\} \log \left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) - \log \rho_0(\Omega_{r_\varepsilon/2}) \right),
 \tag{3.33}$$

where we introduce the definitions

$$c(\vartheta, \chi_1, \lambda_1, \sigma_1) := \min \left\{ \frac{\vartheta}{2} \frac{\chi_1}{2\sqrt{2}(\lambda_1 + \sigma_1^2)}, \sqrt{\frac{\vartheta}{2} \frac{\chi_1}{\sigma_1^2}} \right\}
 \tag{3.34}$$

as well as

$$q_\varepsilon := \frac{1}{2} \min \left\{ \left(\eta \frac{c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\varepsilon}}{2\sqrt{d}} \right)^{1/\nu}, \mathcal{E}_\infty \right\} \quad \text{and} \quad r_\varepsilon := \max_{s \in [0, R_0]} \left\{ \max_{v \in B_s^\infty(x^*)} \mathcal{E}(v) \leq q_\varepsilon \right\}.
 \tag{3.35}$$

Moreover, p is as given in (3.12) in Proposition 3.8 with $B = c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\mathcal{V}(\rho_0)}$ in C_Υ and with $r = r_\varepsilon$. By construction, $q_\varepsilon > 0$ and $r_\varepsilon \leq R_0$. Furthermore, recalling the notation $\mathcal{E}_r = \sup_{v \in B_r^\infty(x^*)} \mathcal{E}(v)$ from Proposition 3.6, we have $q_\varepsilon + \mathcal{E}_{r_\varepsilon} \leq 2q_\varepsilon \leq \mathcal{E}_\infty$ according to the definition of r_ε . Since $q_\varepsilon > 0$, the continuity of \mathcal{E} ensures that there exists $s_{q_\varepsilon} > 0$ such that $\mathcal{E}(v) \leq q_\varepsilon$ for all $v \in B_{s_{q_\varepsilon}}^\infty(x^*)$, yielding also $r_\varepsilon > 0$.

Let us now define the time horizon $T_\alpha \geq 0$ by

$$T_\alpha := \sup \left\{ t \geq 0 : \mathcal{V}(\rho_t) > \varepsilon \text{ and } \|y_\alpha(\rho_{Y,t'}) - x^*\|_2 < C(t') \text{ for all } t' \in [0, t] \right\}
 \tag{3.36}$$

with $C(t) := c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\mathcal{V}(\rho_t)}$. Notice for later use that $C(0) = B$.

Our aim now is to show that $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in \left[\frac{(1-\vartheta)\chi_1}{(1+\vartheta/2)\chi_2} T^*, T^* \right]$ and that we have at least exponential decay of $\mathcal{V}(\rho_t)$ until time T_α , i.e., until the accuracy ε is reached.

First, however, we verify that $T_\alpha > 0$, which is due to the continuity of $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \|y_\alpha(\rho_{Y,t}) - x^*\|_2$ since $\mathcal{V}(\rho_0) > \varepsilon$ and $\|y_\alpha(\rho_{Y,0}) - x^*\|_2 < C(0)$ at time 0. While the former is a consequence of the assumption, the latter follows from Proposition 3.6 with q_ε and r_ε as defined in (3.35), which allows to show that

$$\begin{aligned} \|y_\alpha(\rho_{Y,0}) - x^*\|_2 &\leq \frac{\sqrt{d}(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y,0}(B_{r_\varepsilon}^\infty(x^*))} \int \|y - x^*\|_2 d\rho_{Y,0}(y) \\ &\leq \frac{\sqrt{d}(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y,0}(B_{r_\varepsilon}^\infty(x^*))} \iint \|y - x\|_2 + \|x - x^*\|_2 d\rho_0(x, y) \\ &\leq \frac{c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\varepsilon}}{2} + \frac{2\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y,0}(B_{r_\varepsilon}^\infty(x^*))} \sqrt{\mathcal{V}(\rho_0)} \\ &\leq c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\varepsilon} < c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\mathcal{V}(\rho_0)} = C(0). \end{aligned}$$

The first inequality in the last line holds by the choice of α in (3.33) and by noticing that $\Omega_{r_\varepsilon/2} \subset \mathbb{R}^d \times B_{r_\varepsilon}^\infty(x^*)$ and thus $\rho_0(\Omega_{r_\varepsilon/2}) \leq \rho_{Y,0}(B_{r_\varepsilon}^\infty(x^*))$.

Next, we show that the functional $\mathcal{V}(\rho_t)$ is sandwiched between two exponentially decaying functions with rates $(1 - \vartheta)\chi_1$ and $(1 + \vartheta/2)\chi_2$, respectively. More precisely, we prove that, up to time T_α , $\mathcal{V}(\rho_t)$ decays

- (i) at least exponentially fast (with rate $(1 - \vartheta)\chi_1$), and
- (ii) at most exponentially fast (with rate $(1 + \vartheta/2)\chi_2$).

To obtain (i), recall that Corollary 3.3 provides an upper bound on the time derivative of $\mathcal{V}(\rho_t)$ given by

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -\chi_1 \mathcal{V}(\rho_t) + 2\sqrt{2} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{V}(\rho_t)} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 + \sigma_1^2 \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2 \tag{3.37}$$

with χ_1 as in (2.12a) being strictly positive by assumption. By combining (3.37) and the definition of T_α in (3.36), we have by construction

$$\frac{d}{dt} \mathcal{V}(\rho_t) \leq -(1 - \vartheta)\chi_1 \mathcal{V}(\rho_t) \quad \text{for all } t \in (0, T_\alpha).$$

Analogously, for (ii), by Corollary 3.5, we obtain a lower bound on the time derivative of $\mathcal{V}(\rho_t)$ given by

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(\rho_t) &\geq -\chi_2 \mathcal{V}(\rho_t) - 2\sqrt{2} (\lambda_1 + \sigma_1^2) \sqrt{\mathcal{V}(\rho_t)} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \\ &\geq -(1 + \vartheta/2)\chi_2 \mathcal{V}(\rho_t) \quad \text{for all } t \in (0, T_\alpha), \end{aligned} \tag{3.38}$$

where the second inequality again exploits the definition of T_α . Grönwall’s inequality now implies for all $t \in [0, T_\alpha]$ the upper and lower estimates

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0) \exp(-(1 - \vartheta)\chi_1 t), \tag{3.39a}$$

$$\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0) \exp(-(1 + \vartheta/2)\chi_2 t), \tag{3.39b}$$

thereby proving (i) and (ii). The definition of T_α together with the one of $C(t)$ permits to control

$$\max_{t \in [0, T_\alpha]} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \leq \max_{t \in [0, T_\alpha]} C(t) \leq C(0). \tag{3.40}$$

To conclude, it remains to prove $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in \left[\frac{(1-\vartheta)\chi_1}{(1+\vartheta/2)\chi_2} T^*, T^* \right]$. To this end, we consider the following three cases separately.

Case $T_\alpha \geq T^*$: If $T_\alpha \geq T^*$, the time-evolution bound of $\mathcal{V}(\rho_t)$ from (3.39a) combined with the definition of T^* in (2.13) allows to immediately infer $\mathcal{V}(\rho_{T^*}) \leq \varepsilon$. Therefore, with $\mathcal{V}(\rho_t)$ being continuous, $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ and $T_\alpha = T^*$ according to the definition of T_α in (3.36).

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) \leq \varepsilon$: By continuity of $\mathcal{V}(\rho_t)$, it holds for T_α as defined in (3.36), $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$. Thus, $\varepsilon = \mathcal{V}(\rho_{T_\alpha}) \geq \mathcal{V}(\rho_0) \exp(-(1 + \vartheta/2)\chi_2 T_\alpha)$ as a consequence of the time-evolution bound (3.39b). The latter can be reordered as

$$\frac{(1 - \vartheta)\chi_1}{(1 + \vartheta/2)\chi_2} T^* = \frac{1}{(1 + \vartheta/2)\chi_2} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) \leq T_\alpha < T^*.$$

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$: We will prove that this case can actually not occur by showing that $\|y_\alpha(\rho_{Y, T_\alpha}) - x^*\|_2 < C(T_\alpha)$ for the α chosen in (3.33). In fact, if both $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$ and $\|y_\alpha(\rho_{Y, T_\alpha}) - x^*\|_2 < C(T_\alpha)$ held true simultaneously, this would contradict the definition of T_α in (3.36). To obtain this contradiction, we apply again Proposition 3.6 with q_ε and r_ε as before to get

$$\begin{aligned} \|y_\alpha(\rho_{Y, T_\alpha}) - x^*\|_2 &\leq \frac{\sqrt{d}(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y, T_\alpha}(B_{r_\varepsilon}^\infty(x^*))} \int \|y - x^*\|_2 d\rho_{Y, T_\alpha}(y) \\ &\leq \frac{\sqrt{d}(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y, T_\alpha}(B_{r_\varepsilon}^\infty(x^*))} \iint \|y - x\|_2 + \|x - x^*\|_2 d\rho_{T_\alpha}(x, y) \\ &\leq \frac{c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\varepsilon}}{2} + \frac{2\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y, T_\alpha}(B_{r_\varepsilon}^\infty(x^*))} \sqrt{\mathcal{V}(\rho_{T_\alpha})} \\ &< \frac{c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2\sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_{Y, T_\alpha}(B_{r_\varepsilon}^\infty(x^*))} \sqrt{\mathcal{V}(\rho_{T_\alpha})}. \end{aligned} \tag{3.41}$$

Since, thanks to (3.40), we have $\max_{t \in [0, T_\alpha]} \|y_\alpha(\rho_{Y, t}) - x^*\|_2 \leq B$ for $B = C(0)$, which in particular does not depend on α , Proposition 3.8 guarantees the existence of $p > 0$ independent of α (but dependent on B and r_ε) with

$$\begin{aligned} \rho_{Y, T_\alpha}(B_{r_\varepsilon}^\infty(x^*)) &\geq \left(\iint \phi_{r_\varepsilon}(x, y) d\rho_0(x, y) \right) \exp(-pT_\alpha) \\ &\geq \frac{1}{2^d} \rho_0(\Omega_{r_\varepsilon/2}) \exp(-pT^*) > 0. \end{aligned}$$

Here we use that $(x^*, x^*) \in \text{supp}(\rho_0)$ to bound the initial mass ρ_0 and the fact that ϕ_r from Lemma 3.7 is bounded from below on $\Omega_{r/2}$ by $1/2^d$. With this, we can continue the chain of inequalities in (3.41)3 to obtain

$$\begin{aligned} \|y_\alpha(\rho_{Y, T_\alpha}) - x^*\|_2 &< \frac{c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2^{d+1} \sqrt{d} \exp(-\alpha q_\varepsilon)}{\rho_0(\Omega_{r_\varepsilon/2}) \exp(-pT^*)} \sqrt{\mathcal{V}(\rho_{T_\alpha})} \\ &\leq c(\vartheta, \chi_1, \lambda_1, \sigma_1) \sqrt{\mathcal{V}(\rho_{T_\alpha})} = C(T_\alpha), \end{aligned}$$

with the first inequality in the last line holding due to the choice of α in (3.33). This gives the desired contradiction, again thanks to the continuity of $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \|y_\alpha(\rho_{Y, t}) - x^*\|_2$. \square

4. Numerical experiments

In the first part of this section, we comment on how to efficiently implement a numerical scheme for the CBO dynamics (1.1) which allows to integrate memory mechanisms without additional computational complexity. Afterwards, we numerically demonstrate the benefit of memory effects and gradient information at the example of interesting real-world inspired applications.

4.1. Implementational aspects

Discretising the interacting particle system (1.1) in time by means of the Euler-Maruyama method [37] with prescribed time step size Δt results in the implementable numerical scheme

$$X_{k+1}^i = X_k^i - \Delta t \lambda_1 (X_k^i - y_\alpha(\widehat{\rho}_{Y,k}^N)) - \Delta t \lambda_2 (X_k^i - Y_k^i) - \Delta t \lambda_3 \nabla \mathcal{E}(X_k^i) + \sigma_1 D(X_k^i - y_\alpha(\widehat{\rho}_{Y,k}^N)) B_k^{1,i} + \sigma_2 D(X_k^i - Y_k^i) B_k^{2,i} + \sigma_3 D(\nabla \mathcal{E}(X_k^i)) B_k^{3,i}, \tag{4.1a}$$

$$Y_{k+1}^i = Y_k^i + \Delta t \kappa (X_{k+1}^i - Y_k^i) S^{\beta,\theta}(X_{k+1}^i, Y_k^i), \tag{4.1b}$$

where $((B_k^{m,i})_{k=0,\dots,K-1})_{i=1,\dots,N}$ are independent, identically distributed Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}$ for $m \in \{1, 2, 3\}$.

We notice that, compared to standard CBO, see, e.g., [28, Equation (2)], the way the historical best position is updated in (4.1b) (recall the definition of $S^{\beta,\theta}$ from equation (1.4)) requires one additional evaluation of the objective function per particle in each time step, which raises the computational complexity of the numerical scheme substantially if computing \mathcal{E} is costly and the dominating part. However, for the parameter choices $\kappa = 1/\Delta t$, $\theta = 0$ and $\beta = \infty$, in place of (4.1b), we obtain the update rule

$$Y_{k+1}^i = \begin{cases} X_{k+1}^i, & \text{if } \mathcal{E}(X_{k+1}^i) < \mathcal{E}(Y_k^i), \\ Y_k^i, & \text{else,} \end{cases} \tag{4.2}$$

which is how one expects a memory mechanism to be implemented. This way allows to recycle in time step k the computations made in the previous step and thus leads to no additional computational cost as consequence of using memory effects. The memory consumption, on the other hand, is approximately twice as high as in standard CBO.

4.2. A benchmark problem in optimisation: the Rastrigin function

Let us validate in this section the numerical observation made in Figure 2a in the introduction about the benefit of memory effects. Namely, it has been observed in several prior works that a higher noise level can enhance the success of CBO. To rule out that the improved performance for $\lambda_2 > 0$ in Figure 2a originates solely from the larger present noise as consequence of the additional noise term associated with the memory drift, we replicate in Figure 3 the experiments with the exception of setting $\sigma_2 = 0$. The obtained results confirm that already the usage of memory effects together with a memory drift improves the performance. However, we also notice that an additional noise term further increases the success probability.

4.3. A machine learning example

As a first real-world inspired application, we now investigate the influence of memory mechanisms in a high-dimensional benchmark problem in machine learning, which is well-understood in the literature, namely the training of a shallow and a convolutional NN (CNN) classifier for the MNIST dataset of handwritten digits [47].

The experimental setting is the one of [29, Section 4] with tested architectures as described in Figure 4. While it is not our aim to challenge the state of the art at this task by employing very sophisticated architectures, we demonstrate that CBO is on par with stochastic gradient descent without requiring time-consuming hyperparameter tuning.

To train the learnable parameters θ of the NNs, we minimise the empirical risk $\mathcal{E}(\theta) = \frac{1}{M} \sum_{j=1}^M \ell(f_\theta(x^j), y^j)$, where f_θ denotes the forward pass of the NN and $(x^j, y^j)_{j=1}^M$ the M training samples consisting of image and label. As loss ℓ we choose the categorical crossentropy loss $\ell(\widehat{y}, y) = - \sum_{k=0}^9 y_k \log(\widehat{y}_k)$ with $\widehat{y} = f_\theta(x)$ denoting the output of the NN for a sample (x, y) .

Our implementation is the one of [29, Section 4], which includes concepts from [14] and [26, Section 2.2]. Firstly, mini-batching is employed when evaluating \mathcal{E} and when computing the consensus point y_α ,

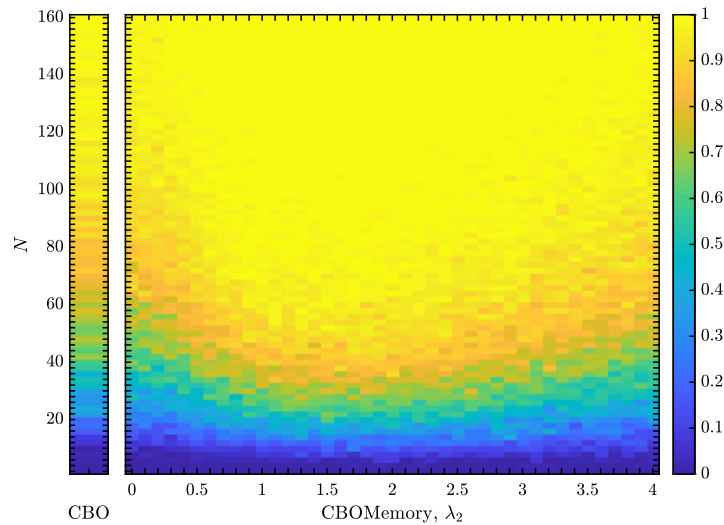


Figure 3. Success probability of CBO without (left separate column) and with memory effects for different values of the parameter $\lambda_2 \in [0, 4]$ (right phase diagram) when optimising the Rastrigin function in dimension $d = 4$ in the setting of Figure 2a with the exception of setting $\sigma_2 = 0$. In this way we validate that the presence of memory effects is responsible for the improved performance and not just a higher noise level.

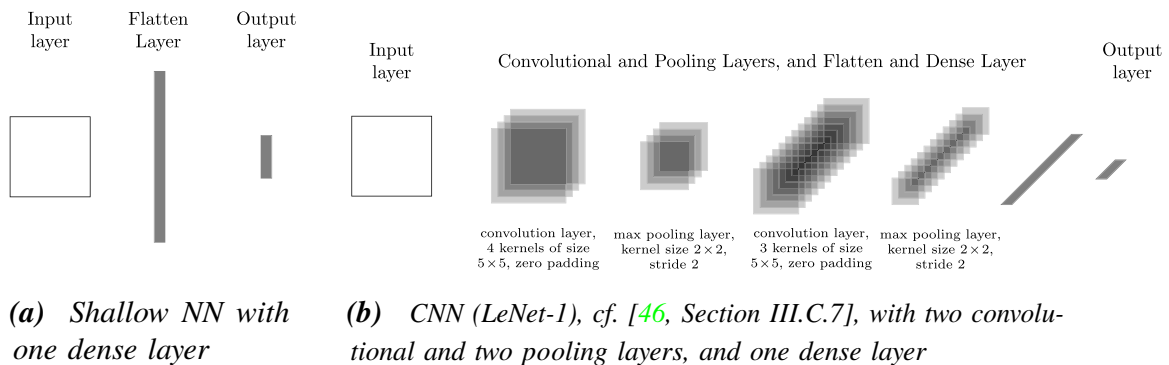
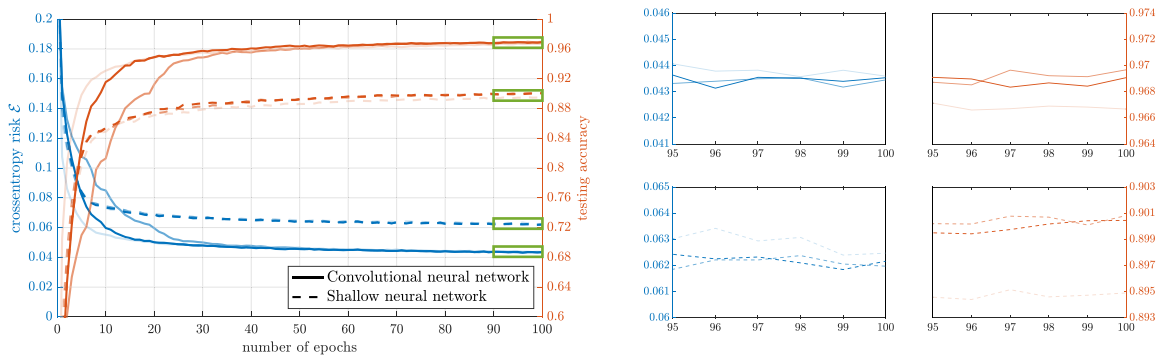


Figure 4. NN architectures used in the experiments of Section 4.3. Images are represented as 28×28 matrices with entries in $[0, 1]$. For the shallow NN in (a) the input is reshaped into a vector $x \in \mathbb{R}^{728}$ which is then passed through a dense layer of the form $\text{ReLU}(Wx + b)$ with trainable weights $W \in \mathbb{R}^{10 \times 728}$ and bias $b \in \mathbb{R}^{10}$. The learnable parameters of the CNN in (b) are the kernels and the final dense layer. Both networks include a batch normalisation step after each ReLU activation function and a softmax activation in the last layer in order to be able to interpret the output as a probability distribution over the digits. We denote the trainable parameters of the NN by θ . The shallow NN has 7850 and the CNN 2112. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Applications of Evolutionary Computation, Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law, M. Fornasier, T. Klock, K. Riedl, © 2022.)

which means that \mathcal{E} is evaluated on a random subset of size $n_{\mathcal{E}} = 60$ of the training dataset and y_{α} is computed from a random subset of size $n_N = 10$ of all $N = 100$ particles. Secondly, a cooling strategy for α and the noise parameters is used. More precisely, α is doubled each epoch, while σ_1 and σ_2 follow the schedule $\sigma_{i,\text{epoch}} = \sigma_{i,0} / \log_2(\text{epoch} + 2)$ for $i = 1, 2$.

In Figure 5, we report the testing accuracies and the training risks evaluated at the consensus point based on a random sample of the training set of size 10,000 for both the shallow NN and the CNN when trained with one of three algorithms: standard CBO without memory effects as obtained when



(a) Testing accuracy and empirical risk plots for the shallow NN and the CNN when trained with CBO without memory effects (lightest lines), with memory effects but without memory drift (line with intermediate opacity) and with memory effects and memory drift (darkest lines)

(b) Zooming into the testing accuracies (right column) and the empirical risks (left column) during the final 5 epochs (highlighted with green boxes in (a)) for the shallow NN (bottom row) and the CNN (top row)

Figure 5. Comparison of the performances (testing accuracy and training loss) of a shallow NN (dashed lines) and a CNN (solid lines) with architectures as described in Figure 4, when trained with CBO without memory effects (lightest lines), with memory effects but without memory drift (line with intermediate opacity) and with memory effects and memory drift (darkest lines). Depicted are the accuracies on a test dataset (orange lines) and the values of the objective function \mathcal{E} evaluated on a random sample of the training set of size 10,000 (blue lines). We observe that memory effects slightly improve the final accuracies while slowing down the training process initially.

discretising [29, Equation (1)], CBO with memory effects but without memory drift as in equation (4.1) with $\lambda_2 = \sigma_2 = 0$, and CBO with memory effects and memory drift as in equation (4.1) with $\lambda_2 = 0.4$ and $\sigma_2 = \lambda_2 \sigma_1$. The remaining parameters are $\lambda_1 = 1, \sigma_{1,0} = \sqrt{0.4}, \alpha_{\text{initial}} = 50, \beta = \infty, \theta = 0, \kappa = 1/\Delta t$, and discrete time step size $\Delta t = 0.1$. We train for 100 epochs and use $N = 100$ particles, which are initialised according to $\mathcal{N}((0, \dots, 0)^T, \text{Id})$. All results are averaged over 5 training runs.

As concluded already in [29, Section 4], we obtain accuracies comparable to SGD, cf. [46, Figure 9]. Moreover, we see slightly improved results when exploiting memory effects. However, we also notice that memory mechanisms slow down the training process initially.

4.4. A compressed sensing example

In the final numerical section of this paper, we showcase an application where gradient information turns out to be indispensable for the success of CBO methods, namely an experiment in compressed sensing [30], which has become a very active and profitable field of research since the seminal works [11, 22] about two decades ago.

One of the most common problems encountered in engineering and technology is concerned with the inference of information about an unknown signal $x^* \in \mathbb{R}^d$ from (linear) measurements $b \in \mathbb{R}^m$. While classical linear algebra suggests that the number of measurements m must be at least as large as the dimensionality d of the signal, in many applications measurements are costly, time-consuming or both, making it desirable to reduce their number to the absolute minimum. Very often one aims at $m \ll d$, since real-world signals usually live in high-dimensional spaces. In general, this would be an impossible task. However, in typical practical scenarios additional information about the quantity of interest x^* is available, which indeed allows to reconstruct signals from few measurements b . An empirically observed assumption about real-world signals is compressibility, meaning that they can be well-approximated by

sparse vectors, i.e., vectors whose components are for the most part zero. Exploiting sparsity enables us to solve the underdetermined system $Ax^* = b$ efficiently in both theory and practice. Compressed sensing is concerned with the task of designing a measurement process $A \in \mathbb{R}^{m \times d}$ together with a reconstruction algorithm capable of recovering the sparse solution x^* from the set of solutions consistent with the measurements. This can be formulated as the nonconvex combinatorial optimisation

$$\min \|x\|_0 \quad \text{subject to } Ax = b,$$

where $\|x\|_0$ is colloquially referred to as ℓ_0 -‘norm’ and denotes the number of non-zero elements of x . Solving ℓ_0 -minimisation is however NP-hard in general, which lead researchers to consider the convex relaxation

$$\min \|x\|_1 \quad \text{subject to } Ax = b. \quad (4.3)$$

ℓ_1 -minimisation is easy to solve by means of established methods from convex optimisation and provably recovers the correct solution for a suitable measurement matrix A . The remaining question is about the correct way of inferring information about the signal through measurements. Remarkably and responsible for the wide success of compressed sensing is that random matrices enjoy properties such as the null space or restricted isometry property, which guarantee successful recovery, for $m \gtrsim s \log(d/s)$, where s denotes the sparsity of the signal x^* , i.e., $s = \|x^*\|_0$. Up to the logarithmic factor in the ambient dimension d , this is optimal, since in theory $m = 2s$ measurements are necessary and sufficient to reconstruct every s -sparse vector.

In the numerical experiments following, we resort to the regularised variant of the sparse recovery problem

$$\min \mathcal{E}(x) \quad \text{with } \mathcal{E}(x) = \frac{1}{2} \|Ax - b\|_2^2 + \mu \|x\|_p^p \quad (4.4)$$

for a suitable regularisation parameter $\mu > 0$. For $p = 1$ we obtain the regularisation of (4.3), whereas for $p < 1$ the optimisation (4.4) becomes nonconvex. Our results in Figures 2b and 6 show that CBO with gradient information is capable of solving the convex but also the nonconvex optimisation problem (4.4) with $p = 1/2$ with already very few measurements. As parameters of the CBO algorithm, which is obtained as a Euler-Maruyama discretisation of equation (1.1), we choose in both cases the time horizon $T = 20$, time step size $\Delta t = 0.01$, $\alpha = 100$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, $\lambda_1 = 1$, $\lambda_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_3 = 0$. We use either $N = 10$ or $N = 100$ particles, which is specified in the respective caption. After running the CBO algorithm, a post-processing step is performed, in which the support of the suspected sparse vector is identified by checking which entries are not smaller than 0.01 before the final sparse solution is then obtained by solving the linear system constrained to this support.

The depicted success probabilities are averaged over 100 runs of CBO. In Figure 2b, we solve the sparse recovery problem in the convex setting for an 8-sparse 200-dimensional signal with $p = 1$ using CBO without and with gradient information with merely 10 particles. We observe that gradient information is indispensable to be able to identify the correct sparse solution and standard CBO would fail in such task. In Figure 6, we conduct a slightly lower-dimensional experiment with a 2-sparse 50-dimensional signal. Here our focus is to enter the nonconvex recovery regime by comparing the convex ℓ_1 -regularised with the nonconvex $\ell_{1/2}$ -regularised problem. We discover that in either case reconstruction is feasible from already very few measurements. Increasing the number of optimising particles has almost no effect for the convex optimisation problem, in the nonconvex setting recovery benefits from more particles. Furthermore, the nonconvex problem demands a more moderate choice of the strength of the gradient.

5. Conclusions

In this paper, we investigate a variant of consensus-based optimisation (CBO) which incorporates memory effects and gradient information. By developing further and generalising the proof technique devised in [28, 29], we establish the global convergence of the underlying dynamics to the global minimiser x^* of the objective function \mathcal{E} in mean-field law. To this end, we analyse the time-evolution of the Wasserstein

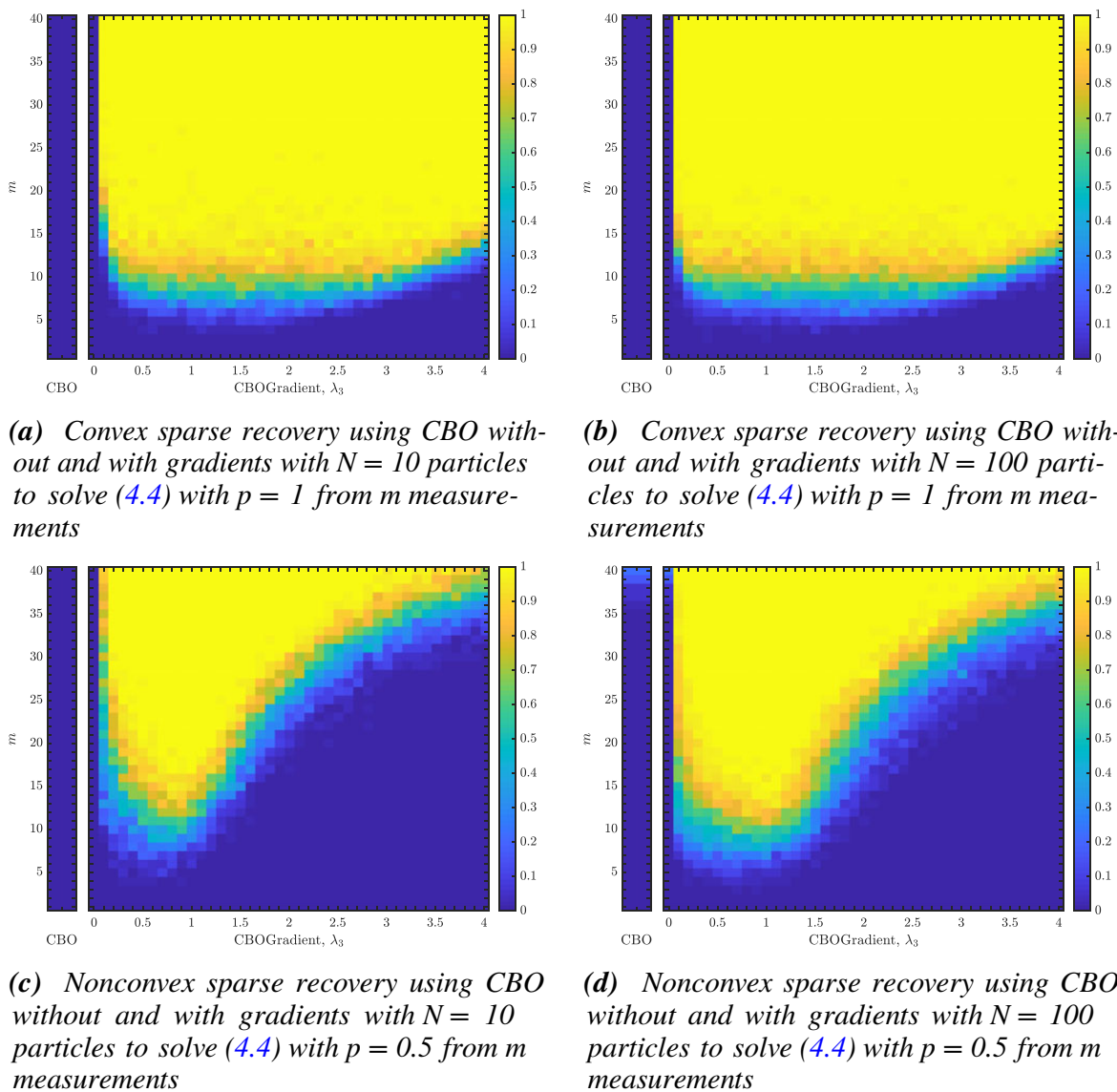


Figure 6. Comparison between the success probabilities of CBO without (left separate columns) and with gradient information for different values of the parameter $\lambda_3 \in [0, 4]$ (right phase diagrams) with $N = 10$ ((a) and (c)) or $N = 100$ particles ((b) and (d)) when solving the convex or nonconvex ℓ_p -regularised least squares problem (4.4) with $p = 1$ and $\mu = ((a) \text{ and } (b))$ or $p = 0.5$ and $\mu = ((c) \text{ and } (d))$, respectively. On the vertical axis we depict the number of measurements m , from which we try to recover the 2-sparse and 50-dimensional sparse signal. As further parameters we choose the time horizon $T = 20$, discrete time step size $\Delta t = 0.01$, $\alpha = 100$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, $\lambda_1 = 1$, $\lambda_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_3 = 0$. We discover that in both the convex and nonconvex setting reconstruction is feasible from already very few measurements. While increasing the number of optimising particles has almost no effect for the convex optimisation problem, in the nonconvex setting recovery benefits from more particles. Note that the separate columns and the left most column of the phase diagrams coincide, and are only depicted in this way to highlight that we compare also CBO.

distance between the law of the mean-field CBO dynamics and a Dirac delta at the minimiser and show its exponential decay in time. Our result holds under minimal assumptions about the initial measure ρ_0 and for a vast class of objective functions. The numerical benefit of such additional terms, specifically the employed memory effects and gradient information, is demonstrated at the example of a benchmark function in optimisation as well as at real-world applications such as compressed sensing and the training of neural networks for image classification.

By these means, we demonstrate the versatility, flexibility and customisability of the class of CBO methods, both with respect to potential application areas in practice and modifications to the underlying optimisation principles, while still being amenable to theoretical analysis.

Acknowledgements. I would like to profusely thank Massimo Fornasier for many fruitful and stimulating discussions while I was preparing this manuscript.

Financial support. This work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. I, the author of this work, take full responsibility for its content. Furthermore, I acknowledge the financial support from the Technical University of Munich – Institute for Ethics in Artificial Intelligence. Moreover, I gratefully acknowledge the computer and data resources provided by the Leibniz Supercomputing Centre.

Competing interests. None.

References

- [1] Albi, G., Bongini, M., Cristiani, E. & Kalise, D. (2015) Invisible sparse control of self-organizing agents leaving unknown environments. *SIAM J. Appl. Math.* **76**(4), 1683–1710.
- [2] Ambrosio, L., Gigli, N. & Savaré, G. (2008) *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics ETH Zürich, 2nd ed., Birkhäuser Verlag, Basel.
- [3] Arnold, L. (1974) *Stochastic Differential Equations: Theory and Applications*, Wiley-Interscience [John Wiley & Sons], New York/London/Sydney. Translated from the German.
- [4] Back, T., Fogel, D. B. & Michalewicz, Z. (1997) *Handbook of Evolutionary Computation*, Institute of Physics Publishing, Bristol; Oxford University Press, New York.
- [5] Bae, H.-O., Ha, S.-Y., Kang, M., Lim, H., Min, C. & Yoo, J. (2022) A constrained consensus based optimization algorithm and its application to finance. *Appl. Math. Comput.* **416**, PaperNo.126726,10.
- [6] Blum, C. & Roli, A. (2003) Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Comput. Surv.* **35**(3), 268–308.
- [7] Borghi, G., Herty, M. & Pareschi, L. (2022) A consensus-based algorithm for multi-objective optimization and its mean-field description. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, pp. 4131–4136.
- [8] Borghi, G., Herty, M. & Pareschi, L. (2023) An adaptive consensus based method for multi-objective optimization with uniform Pareto front approximation. *Appl. Math. Optim.* **88**(2), 1–43.
- [9] Borghi, G., Herty, M. & Pareschi, L. (2023) Constrained consensus-based optimization. *SIAM J. Optim.* **33**(1), 211–236.
- [10] Bungert, L., Wacker, P. & Roith, T. (2022) Polarized consensus-based dynamics for optimization and sampling. *arXiv: 2211.05238*.
- [11] Candès, E. J., Romberg, J. K. & Tao, T. (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**(8), 1207–1223.
- [12] Carrillo, J. A., Choi, Y.-P., Totzeck, C. & Tse, O. (2018) An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.* **28**(6), 1037–1066.
- [13] Carrillo, J. A., Hoffmann, F., Stuart, A. M. & Vaes, U. (2022) Consensus-based sampling. *Stud. Appl. Math.* **148**(3), 1069–1140.
- [14] Carrillo, J. A., Jin, S., Li, L. & Zhu, Y. (2021) A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.* **27**, Paper No.S5,22.
- [15] Carrillo, J. A., Totzeck, C. & Vaes, U. (2023) Consensus-based optimization and ensemble Kalman inversion for global optimization problems with constraints. In: *Modeling and Simulation for Collective Dynamics*, World Scientific, Singapore, pp. 195–230.
- [16] Carrillo, J. A., Trillos, N. G., Li, S. & Zhu, Y. (2023). FedCBO: reaching group consensus in clustered federated learning through consensus-based optimization. *arXiv: 2305.02894*.
- [17] Chen, J., Jin, S. & Lyu, L. (2020) A consensus-based global optimization method with adaptive momentum estimation. *arXiv: 2012.04827*.
- [18] Cipriani, C., Huang, H. & Qiu, J. (2022) Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Appl. Math.* **54**(3), 3091–3121.
- [19] Couzin, I. D., Krause, J., Franks, N. R. & Levin, S. A. (2005) Effective leadership and decision-making in animal groups on the move. *Nature* **433**(7025), 513–516.
- [20] Cristiani, E., Piccoli, B. & Tosin, A. (2011) Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Model. Simul.* **9**(1), 155–182.
- [21] Dembo, A. & Zeitouni, O. (1998) Large Deviations Techniques and Applications, *Applications of Mathematics (New York)*, Vol. **38**, 2nd ed., Springer-Verlag, New York.
- [22] Donoho, D. L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306.
- [23] Dorigo, M. & Blum, C. (2005) Ant colony optimization theory: a survey. *Theor. Comput. Sci.* **344**(2-3), 243–278.
- [24] Fogel, D. B. (2000) *Evolutionary Computation. Toward a New Philosophy of Machine Intelligence*, 2nd ed., IEEE Press, Piscataway, NJ.

- [25] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2020) Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.* **30**(14), 2725–2751.
- [26] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2021) Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.* **22**, PaperNo.237,55.
- [27] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2022) Anisotropic diffusion in consensus-based optimization on the sphere. *SIAM J. Optim.* **32**(3), 1984–2012.
- [28] Fornasier, M., Klock, T. & Riedl, K. (2021). Consensus-based optimization methods converge globally. *arXiv: 2103.15130*.
- [29] Fornasier, M., Klock, T. & Riedl, K. (2022) Convergence of anisotropic consensus-based optimization in mean-field law. In: Jiménez Laredo, J. L., Hidalgo, J. I. & Babaagba, K. O. (editors), *Applications of Evolutionary Computation*, Springer International Publishing, Cham, pp. 738–754.
- [30] Foucart, S. & Rauhut, H. (2013) A mathematical introduction to compressive sensing. In: *Applied and Numerical Harmonic Analysis*, Birkhäuser/Springer, New York.
- [31] Grassi, S., Huang, H., Pareschi, L. & Qiu, J. (2023). Mean-field particle swarm optimization. In: *Modeling and Simulation for Collective Dynamics* (pp. 127–193).
- [32] Grassi, S. & Pareschi, L. (2021) From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Math. Models Methods Appl. Sci.* **31**(8), 1625–1657.
- [33] Ha, S.-Y., Jin, S. & Kim, D. (2020) Convergence of a first-order consensus-based global optimization algorithm. *Math. Models Methods Appl. Sci.* **30**(12), 2417–2444.
- [34] Ha, S.-Y., Jin, S. & Kim, D. (2021) Convergence and error estimates for time-discrete consensus-based optimization algorithms. *Numer. Math.* **147**(2), 255–282.
- [35] Ha, S.-Y., Kang, M. & Kim, D. (2022) Emergent behaviors of high-dimensional Kuramoto models on Stiefel manifolds. *Automatica* **136**, PaperNo.110072.
- [36] Hegselmann, R. & Krause, U. (2002) Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simul.* **5**(3).
- [37] Higham, D. J. (2001) An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.* **43**(3), 525–546.
- [38] Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, Ann Arbor, MI.
- [39] Huang, H. & Qiu, J. (2022) On the mean-field limit for the consensus-based optimization. *Math. Methods Appl. Sci.* **45**(12), 7814–7831.
- [40] Huang, H., Qiu, J. & Riedl, K. (2022) Consensus-based optimization for saddle point problems. *arXiv:2212.12334*.
- [41] Huang, H., Qiu, J. & Riedl, K. (2023) On the global convergence of particle swarm optimization methods. *Appl. Math. Optim.* **88**(2), 30.
- [42] Kalise, D., Sharma, A. & Tretyakov, M. V. (2023) Consensus-based optimization via jump-diffusion stochastic differential equations. *Math. Models Methods Appl. Sci.* **33**(02), 289–339.
- [43] Kennedy, J. & Eberhart, R. (1995) Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*, Vol. 4, IEEE, pp. 1942–1948.
- [44] Kim, J., Kang, M., Kim, D., Ha, S.-Y. & Yang, I. (2020). A stochastic consensus method for nonconvex optimization on the Stiefel manifold. In: *2020 59th IEEE Conference on Decision and Control (CDC)*, IEEE, pp. 1050–1057.
- [45] Klamroth, K., Stiglmayr, M. & Totzeck, C. (2022) Consensus-based optimization for multi-objective problems: a multi-swarm approach. *arXiv: 2211.15737*.
- [46] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324.
- [47] LeCun, Y., Cortes, C. & Burges, C. (2010) MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>
- [48] Miller, P. D. (2006) *Applied Asymptotic Analysis*, Graduate Studies in Mathematics, Vol. 75, American Mathematical Society, Providence, RI.
- [49] Moulines, E. & Bach, F. (2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: *Advances in Neural Information Processing Systems 24*.
- [50] Øksendal, B. (2003) *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer-Verlag, Berlin.
- [51] Parrish, J. K. & Edelstein-Keshet, L. (1999) Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science* **284**(5411), 99–101.
- [52] Pinnau, R., Totzeck, C., Tse, O. & Martin, S. (2017) A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.* **27**(1), 183–204.
- [53] Riedl, K., Klock, T., Geldhauser, C. & Fornasier, M. (2023) Gradient is all you need? *arXiv: 2306.09778*.
- [54] Schillings, C., Totzeck, C. & Wacker, P. (2023) Ensemble-based gradient inference for particle methods in optimization and sampling. *SIAM/ASA J. Uncertain. Quantif.* **11**(3), 757–787.
- [55] Totzeck, C. & Wolfram, M.-T. (2020) Consensus-based global optimization with personal best. *Math. Biosci. Eng.* **17**(5), 6026–6044.
- [56] Vicsek, T. & Zafeiris, A. (2012) Collective motion. *Phys. Rep.* **517**(3-4), 71–140.

Cite this article: Riedl K. Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law. *European Journal of Applied Mathematics*, <https://doi.org/10.1017/S0956792523000293>

License for [CBO-IV].

The permission to reprint and include the material is printed on the next page(s).



RightsLink



Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law

**Author:**

Konstantin Riedl©right=© The Author(s), 2023. Published by Cambridge University Press

Publication: European Journal of Applied Mathematics**Publisher:** Cambridge University Press**Date:** Oct 20, 2023*Copyright © 2023, Cambridge University Press*

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

© 2024 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Data Security and Privacy](#)
| [For California Residents](#) | [Terms and Conditions](#)Comments? We would like to hear from you. E-mail us at customer care@copyright.com

Paper P5

CBX: Python and Julia Packages for Consensus-Based Interacting Particle Methods

R. Bailo, A. Barbaro, S. N. Gomes, K. Riedl, T. Roith, C. Totzeck, and U.
Vaes
arXiv preprint arXiv:2403.14470 (2024)

Paper Summary of [CBX]³⁸

In the paper “CBX: Python and Julia packages for consensus-based interacting particle methods,” we give an overview of the packages [CBXpy](#) and [CBX.jl](#), which provide unified Python and Julia implementations of several consensus-based interacting particle methods.

CBO, as originally proposed by the authors of [Pin+17], is a multi-particle meta-heuristic derivative-free optimization method suitable for tackling nonconvex nonsmooth optimization problems of the form (2.1). Since its introduction, several variants of the algorithm have been proposed and analyzed.

With the packages [CBXpy](#) and [CBX.jl](#) we offer a lightweight, easy-to-understand, -use and -extend implementation of CBO together with several of those variants, including CBO with mini-batch ideas [Car+21; CBO-II], CBO with restart [Car+21; CBO-II], a cooling strategy of the parameters [For+21; CBO-II], polarized CBO [BWR22], CBO with memory effects [GP21; CBO-IV], PSO [GP21; Gra+23; PSO], CBS [Car+22], and more to come. The zoo of different variants of CBO coined the acronym CBX. The defined structures and hierarchies in the code ensure a usage experience similar to optimizer classes in [scikit-opt](#) and [PyTorch](#) [Pas+19]. Numerous utilities, like performance evaluation or plotting routines tailored to CBO methods are provided. The code of these packages builds upon the repositories [polarcbo](#), where polarized CBO [BWR22] is implemented, as well as [cbo-in-python](#), and [Consensus.jl](#), respectively.

KR’s Contributions. The wish to create Python and Julia packages for CBO methods and its variants, which eventually led to the present work, started at the Lorentz Center in Leiden during the workshop “Purpose-driven particle systems” in Spring 2023, which was attended by several of the authors. Already before that, KR supervised the Master’s student project of Igor Tukh that resulted in the Python package [cbo-in-python](#), which had considerable influence on the present [CBXpy](#) package. Independently and at a similar time, TR wrote the Python package [polarcbo](#) for the polarized CBO variant, which eventually laid the conceptual and structural foundation of the present [CBXpy](#) package. Again, independently and at that time, RB wrote the Julia package [Consensus.jl](#), which is the foundation of the present [CBX.jl](#) package. In order to come up with unified Python and Julia implementations of CBO and its variants, which share the core functionalities while still having idiomatic code, all authors discussed in joint meetings the conceptualization and structuring of the implementation. The largest and foundational parts of the current [CBXpy](#) code were written by TR. KR contributed the implementation of some variants of CBO, such as CBO with memory effects and PSO. Moreover, KR and TR discussed several implementational details and features that have been used in the CBO literature. Most parts of the manuscript were written by TR, KR, CT, and RB before the paper was discussed, refined, finalized and proofread together by all authors.

³⁸In this section, we follow [CBX].

The following document is a reprint of

[CBX] R. Bailo, A. Barbaro, S. N. Gomes, K. Riedl, T. Roith, C. Totzeck, and U. Vaes. “CBX: Python and Julia packages for consensus-based interacting particle methods.” In: *arXiv preprint arXiv:2403.14470* (2024).

The permission to reprint and include the material is provided after the reprint.

CBX: Python and Julia packages for consensus-based interacting particle methods

Rafael Bailo^{1,*}, Alethea Barbaro², Susana N. Gomes³, Konstantin Riedl^{4,5},
Tim Roith^{6,*}, Claudia Totzeck⁷, and Urbain Vaes^{8,9}

¹Mathematical Institute, University of Oxford

²Technische Universiteit Delft

³Mathematics Institute, University of Warwick

⁴Technical University of Munich

⁵Munich Center for Machine Learning

⁶Helmholtz Imaging, Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

⁷University of Wuppertal

⁸MATERIALS team, Inria Paris

⁹École des Ponts

*Corresponding authors: bailo@maths.ox.ac.uk, tim.roith@desy.de

Summary

We introduce `CBXPy` and `ConsensusBasedX.jl`, Python and Julia implementations of consensus-based interacting particle systems (CBX), which generalise consensus-based optimization methods (CBO) for global, derivative-free optimisation. The *raison d'être* of our libraries is twofold: on the one hand, to offer high-performance implementations of CBX methods that the community can use directly, while on the other, providing a general interface that can accommodate and be extended to further variations of the CBX family. Python and Julia were selected as the leading high-level languages in terms of usage and performance, as well as their popularity among the scientific computing community. Both libraries have been developed with a common *ethos*, ensuring a similar API and core functionality, while leveraging the strengths of each language and writing idiomatic code.

Mathematical background

Consensus-based optimisation (CBO) is an approach to solve, for a given (continuous) *objective function* $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the *global minimisation problem*

$$x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} f(x),$$

i.e., the task of finding the point x^* where f attains its lowest value. Such problems arise in a variety of disciplines including engineering, where x might represent a vector of design parameters for a structure and f a function related to its cost and structural integrity, or machine learning, where x could comprise the parameters of a neural network and f the empirical loss, which measures the discrepancy of the neural network prediction with the observed data.

In some cases, so-called *gradient-based methods* (those that involve updating a guess of x^* by evaluating the gradient ∇f) achieve state-of-the-art performance in the global minimisation problem. However, in scenarios where f is *non-convex* (when f could have many *local minima*), where ∇f is not well-defined, or where the evaluation of ∇f is impractical due to cost or complexity, *derivative-free* methods are needed. Numerous techniques exist for derivative-free optimisation, such as *random* or *pattern search* [22, 43, 27], *Bayesian optimisation* [38] or *simulated annealing* [26]. Here, we focus on *particle-based methods*, specifically, consensus-based optimisation (CBO), as proposed by Pinnau et al. [42], and the consensus-based taxonomy of related techniques, which we term *CBX*.

CBO uses a finite number N of *agents* (particles), $x_t = (x_t^1, \dots, x_t^N)$, to explore the landscape of f without evaluating any of its derivatives (as do other CBX methods). At each time t , the agents evaluate the objective

function at their position, $f(x_t^i)$, and define a *consensus point* c_α . This point is an approximation of the global minimiser x^* , and is constructed by weighing each agent’s position against the *Gibbs-like distribution* $\exp(-\alpha f)$ [7]. More rigorously,

$$c_\alpha(x_t) = \frac{1}{\sum_{i=1}^N \omega_\alpha(x_t^i)} \sum_{i=1}^N x_t^i \omega_\alpha(x_t^i), \quad \text{where } \omega_\alpha(\cdot) = \exp(-\alpha f(\cdot)),$$

for some $\alpha > 0$. The exponential weights in the definition favour those points x_t^i where $f(x_t^i)$ is lowest, and comparatively ignore the rest, particularly for larger α . If all the found values of the objective function are approximately the same, $c_\alpha(x_t)$ is roughly an arithmetic mean. Instead, if one particle is much better than the rest, $c_\alpha(x_t)$ will be very close to its position.

Once the consensus point is computed, the particles evolve in time following the *stochastic differential equation* (SDE)

$$dx_t^i = -\lambda \underbrace{(x_t^i - c_\alpha(x_t))}_{\text{consensus drift}} dt + \sigma \underbrace{\|x_t^i - c_\alpha(x_t)\|}_{\text{scaled diffusion}} dB_t^i,$$

where λ and σ are positive parameters, and where B_t^i are independent Brownian motions in d dimensions. The *consensus drift* is a deterministic term that drives each agent towards the consensus point, with rate λ . Meanwhile, the *scaled diffusion* is a stochastic term that encourages exploration of the landscape. While both the agents’ positions and the consensus point evolve in time, it has been proven that all agents eventually reach the same position and that the consensus point $c_\alpha(x_t)$ is a good approximation of x^* [11, 18]. Other variations of the method, such as CBO with anisotropic noise [14], *polarised CBO* [10], or *consensus-based sampling* (CBS) [13] have also been proposed.

In practice, the solution to the SDE above cannot be found exactly. Instead, an *Euler–Maruyama scheme* [35] is used to update the position of the agents. The update is given by

$$x^i \leftarrow x^i - \lambda \Delta t (x^i - c_\alpha(x)) + \sigma \sqrt{\Delta t} \|x^i - c_\alpha(x)\| \xi^i,$$

where $\Delta t > 0$ is the *step size* and $\xi^i \sim \mathcal{N}(0, \text{Id})$ are independent, identically distributed, standard normal random vectors.

As a particle-based family of methods, CBX is conceptually related to other optimisation approaches which take inspiration from biology, like *particle-swarm optimisation* (PSO) [32], from physics, like *simulated annealing* (SA) [26], or from other heuristics [40, 31, 48, 4]. However, unlike many such methods, CBX has been designed to be compatible with rigorous convergence analysis at the mean-field level (the infinite-particle limit, see [28]). Many convergence results have been shown, whether in the original formulation [11, 18], for CBO with anisotropic noise [14, 19], with memory effects [44], with truncated noise [20], for polarised CBO [10], and PSO [30]. The relation between CBO and *stochastic gradient descent* has been recently established by Riedl et al. [45], which suggests a previously unknown yet fundamental connection between derivative-free and gradient-based approaches.

CBX methods have been successfully applied and extended to several different settings, such as constrained optimisation problems [17, 9], multi-objective optimisation [8, 34], saddle-point problems [29], federated learning tasks [12], uncertainty quantification [2], or sampling [13].

Statement of need

In general, very few implementations of CBO already exist, and none have been designed with the generality of other CBX methods in mind. We summarise here the related software:

Regarding Python, we refer to Duan et al. [16] and Guo [24] for a collection of various derivative-free optimisation strategies. A very recent implementation of Bayesian optimisation is described by Kim and Choi [33]. PSO and SA implementations are already available [37, 24, 21, 6]. They are widely used by the community and provide a rich framework for the respective methods. However, adjusting these implementations to CBO is not straightforward. The first publicly available Python packages implementing CBX algorithms were given by some of the authors together with collaborators. Tikh and Riedl [47] implement

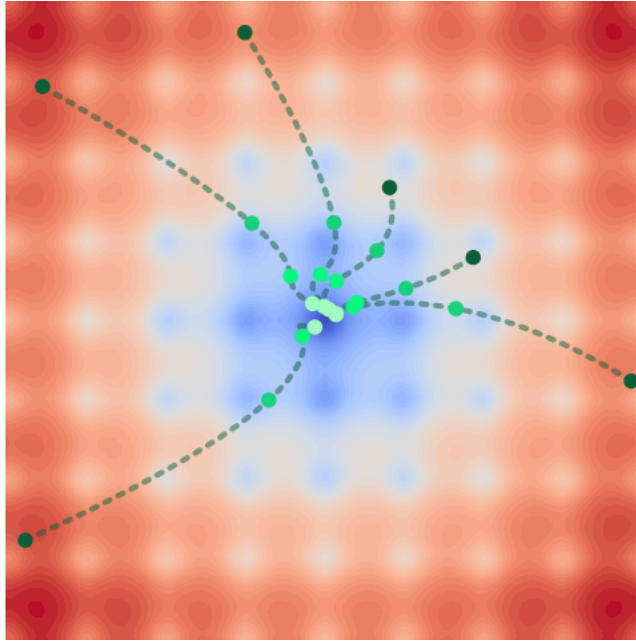


Figure 1: Typical evolution of a CBO method minimising the Ackley function [1].

standard CBO [42], and Roith, Bungert, and Wacker [46] provide an implementation of polarised CBO [10]. [CBXPy](#) is a significant extension of the latter.

Regarding Julia, PSO and SA methods are, among others, implemented by Mogensen and Riseth [39], Mejía-de-Dios and Mezura-Montes [36], and Bergmann [5]. PSO and SA are also included in the meta-library [15], as well as Nelder–Mead, which is a direct search method. One of the authors gave the first specific Julia implementation of standard CBO [3]; that package has now been deprecated in favour of [ConsensusBasedX.jl](#), which offers additional CBX methods and a far more general interface.

Features

[CBXPy](#) and [ConsensusBasedX.jl](#) provide a lightweight and easy-to-understand high-level interface. An existing function can be optimised with just one call. Method selection, parameters, different approaches to particle initialisation, and termination criteria can be specified directly through this interface, offering a flexible point of entry for the casual user. Some of the methods provided are standard CBO [42], CBO with mini-batching [14], polarised CBO [10], CBO with memory effects [23, 44], and consensus-based sampling (CBS) [13]. Parallelisation tools are available.

A more proficient user will benefit from the fully documented interface, which allows the specification of advanced options (e.g., debug output, the noise model, or the numerical approach to the matrix square root of the covariance matrix). Both libraries offer performance evaluation methods as well as visualisation tools.

Ultimately, a low-level interface (including documentation and full-code examples) is provided. Both libraries have been designed to express common abstractions in the CBX family while allowing customisation. Users can easily implement new CBX methods or modify the behaviour of the existing implementation by strategically overriding certain hooks. The stepping of the methods can also be controlled manually.

CBXPy specifics

Most of the [CBXPy](#) implementation uses basic Python functionality, and the agents are handled as an array-like structure. For certain specific features, like broadcasting-behaviour, array copying, and index selection, we fall back to the [numpy](#) implementation [25]. However, it should be noted that an adaptation to other array



Figure 2: CBXPy logo.

or tensor libraries like PyTorch [41] is straightforward. Compatibility with the latter enables gradient-free deep learning directly on the GPU, as demonstrated in the documentation.

The library is available on [GitHub](#) and can be installed via `pip`. It is licensed under the MIT license. The [documentation](#) is available online.

ConsensusBasedX.jl specifics



Figure 3: ConsensusBasedX.jl logo.

[ConsensusBasedX.jl](#) has been almost entirely written in native Julia (with the exception of a single call to LAPACK). The code has been developed with performance in mind, thus the critical routines are fully type-stable and allocation-free. A specific tool is provided to benchmark a typical method iteration, which can be used to detect allocations. Through this tool, unit tests are in place to ensure zero allocations in all the provided methods. The benchmarking tool is also available to users, who can use it to test their implementations of f , as well as any new CBX methods.

The library is available on [GitHub](#). It has been registered in the [general Julia registry](#), and therefore it can be installed by running `Jadd ConsensusBasedX`. It is licensed under the MIT license. The [documentation](#) is available online.

Acknowledgements

We thank the Lorentz Center in Leiden for their kind hospitality during the workshop “Purpose-driven particle systems” in Spring 2023, where this work was initiated. RB was supported by the Advanced Grant Nonlocal-CPD (Nonlocal PDEs for Complex Particle Dynamics: Phase Transitions, Patterns and Synchronisation) of the European Research Council Executive Agency (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 883363) and by the EPSRC grant EP/T022132/1 “Spectral element methods for fractional differential equations, with applications in applied analysis and medical imaging”. KR acknowledges support from the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. TR acknowledges support from DESY (Hamburg, Germany), a member of the Helmholtz Association HGF. This research was supported in part through the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. UV acknowledges support from the Agence Nationale de la Recherche under grant ANR-23-CE40-0027 (IPSO).

References

- [1] David Ackley. *A connectionist machine for genetic hillclimbing*. Vol. 28. Springer science & business media, 2012.
- [2] Konstantin Althaus, Iason Papaioannou, and Elisabeth Ullmann. “Consensus-based rare event estimation”. In: *Preprint arXiv:2304.09077* (2023).

-
- [3] Rafael Bailo. *Consensus.jl*. Version 1.0.0. 2023. DOI: [10.5281/zenodo.7754236](https://doi.org/10.5281/zenodo.7754236). URL: <https://github.com/rafaelbailo/Consensus.jl>.
- [4] Zikri Bayraktar, Muge Komurcu, Jeremy A Bossard, and Douglas H Werner. “The wind driven optimization technique and its application in electromagnetics”. In: *IEEE transactions on antennas and propagation* 61.5 (2013), pp. 2745–2757.
- [5] Ronny Bergmann. “Manopt.jl: Optimization on Manifolds in Julia”. In: *Journal of Open Source Software* 7.70 (2022), p. 3866. DOI: [10.21105/joss.03866](https://doi.org/10.21105/joss.03866).
- [6] Francesco Biscani, Dario Izzo, and Marcus Märtens. *esa/pagmo2: pagmo 2.6*. Nov. 2017. DOI: [10.5281/zenodo.1054110](https://doi.org/10.5281/zenodo.1054110).
- [7] Ludwig Boltzmann. “Studien ueber das Gleichgewicht der lebenden Kraft”. In: *Wissenschaftliche Abhandlungen* 1 (1868), pp. 49–96.
- [8] Giacomo Borghi, Michael Herty, and Lorenzo Pareschi. “An adaptive consensus based method for multi-objective optimization with uniform Pareto front approximation”. In: *Applied Mathematics & Optimization* 88.2 (2023), pp. 1–43.
- [9] Giacomo Borghi, Michael Herty, and Lorenzo Pareschi. “Constrained consensus-based optimization”. In: *SIAM J. Optim.* 33.1 (2023), pp. 211–236.
- [10] Leon Bungert, Philipp Wacker, and Tim Roith. “Polarized consensus-based dynamics for optimization and sampling”. In: *Preprint arXiv:2211.05238* (2022).
- [11] José A Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. “An analytical framework for consensus-based global optimization method”. In: *Math. Models Methods Appl. Sci.* 28.6 (2018), pp. 1037–1066. ISSN: 0218-2025. DOI: [10.1142/S0218202518500276](https://doi.org/10.1142/S0218202518500276).
- [12] José A Carrillo, Nicolas Garcia Trillos, Sixu Li, and Yuhua Zhu. “FedCBO: Reaching Group Consensus in Clustered Federated Learning through Consensus-based Optimization”. In: *Preprint arXiv:2305.02894* (2023).
- [13] José A Carrillo, Franca Hoffmann, Andrew M Stuart, and Urbain Vaes. “Consensus-based sampling”. In: *Stud. Appl. Math.* 148.3 (2022), pp. 1069–1140.
- [14] José A Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. “A consensus-based global optimization method for high dimensional machine learning problems”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 27 (2021), S5.
- [15] Vaibhav Kumar Dixit and Christopher Rackauckas. *Optimization.jl: A Unified Optimization Package*. 2023. DOI: [10.5281/ZENODO.7738525](https://doi.org/10.5281/ZENODO.7738525).
- [16] Qiqi Duan, Guochen Zhou, Chang Shao, Zhuowei Wang, Mingyang Feng, Yijun Yang, Qi Zhao, and Yuhui Shi. “PyPop7: A Pure-Python Library for Population-Based Black-Box Optimization”. In: *Preprint arXiv:2212.05652* (2022).
- [17] Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. “Consensus-based optimization on the sphere: convergence to global minimizers and machine learning”. In: *J. Mach. Learn. Res.* 22 (2021), Paper No. 237, 55. ISSN: 1532-4435.
- [18] Massimo Fornasier, Timo Klock, and Konstantin Riedl. “Consensus-based optimization methods converge globally”. In: *Preprint arXiv:2103.15130* (2021).
- [19] Massimo Fornasier, Timo Klock, and Konstantin Riedl. “Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law”. In: *Applications of Evolutionary Computation*. Ed. by Juan Luis Jiménez Laredo, J. Ignacio Hidalgo, and Kehinde Oluwatoyin Babaagba. Cham: Springer, 2022, pp. 738–754. ISBN: 978-3-031-02462-7.
- [20] Massimo Fornasier, Peter Richtárik, Konstantin Riedl, and Lukang Sun. “Consensus-Based Optimization with Truncated Noise”. In: *Preprint arXiv:2310.16610* (2023).
- [21] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. “DEAP: Evolutionary Algorithms Made Easy”. In: *J. Mach. Learn. Res.* 13 (July 2012), pp. 2171–2175.

-
- [22] Milton Friedman and Lo J Savage. “Planning experiments seeking maxima”. In: *Techniques of statistical analysis* (1947), pp. 365–372.
- [23] Sara Grassi and Lorenzo Pareschi. “From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit”. In: *Math. Models Methods Appl. Sci.* 31.8 (2021), pp. 1625–1657. ISSN: 0218-2025. DOI: [10.1142/S0218202521500342](https://doi.org/10.1142/S0218202521500342).
- [24] Fei Guo. *scikit-opt*. 2021. URL: <https://github.com/guofei9987/scikit-opt>.
- [25] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2).
- [26] Darrall Henderson, Sheldon H Jacobson, and Alan W Johnson. “The theory and practice of simulated annealing”. In: *Handbook of metaheuristics* (2003), pp. 287–319.
- [27] Robert Hooke and Terry A Jeeves. ““Direct Search” Solution of Numerical and Statistical Problems”. In: *Journal of the ACM (JACM)* 8.2 (1961), pp. 212–229.
- [28] Hui Huang and Jinniao Qiu. “On the mean-field limit for the consensus-based optimization”. In: *Math. Methods Appl. Sci.* 45.12 (2022), pp. 7814–7831. ISSN: 0170-4214.
- [29] Hui Huang, Jinniao Qiu, and Konstantin Riedl. “Consensus-based optimization for saddle point problems”. In: *Preprint arXiv:2212.12334* (2022).
- [30] Hui Huang, Jinniao Qiu, and Konstantin Riedl. “On the global convergence of particle swarm optimization methods”. In: *Applied Mathematics & Optimization* 88.2 (2023), No. 30.
- [31] Dervis Karaboga, Beyza Gorkemli, Celal Ozturk, and Nurhan Karaboga. “A comprehensive survey: artificial bee colony (ABC) algorithm and applications”. In: *Artificial intelligence review* 42 (2014), pp. 21–57.
- [32] James Kennedy and Russell Eberhart. “Particle swarm optimization”. In: *Proceedings of ICNN’95-international conference on neural networks*. Vol. 4. IEEE. 1995, pp. 1942–1948.
- [33] Jungtaek Kim and Seungjin Choi. “BayesO: A Bayesian optimization framework in Python”. In: *Journal of Open Source Software* 8.90 (2023), p. 5320. DOI: [10.21105/joss.05320](https://doi.org/10.21105/joss.05320).
- [34] Kathrin Klamroth, Michael Stiglmayr, and Claudia Totzeck. “Consensus-based optimization for multi-objective problems: a multi-swarm approach”. In: *J. Global Optim.* (Feb. 2024). ISSN: 1573-2916. DOI: [10.1007/s10898-024-01369-1](https://doi.org/10.1007/s10898-024-01369-1).
- [35] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin Heidelberg, 1992. ISBN: 9783662126165. DOI: [10.1007/978-3-662-12616-5](https://doi.org/10.1007/978-3-662-12616-5).
- [36] Jesús-Adolfo Mejía-de-Dios and Efrén Mezura-Montes. “Metaheuristics: A Julia package for single-and multi-objective optimization”. In: *Journal of Open Source Software* 7.78 (2022), p. 4723.
- [37] Lester James Miranda. “PySwarms: a research toolkit for Particle Swarm Optimization in Python”. In: *Journal of Open Source Software* 3.21 (2018), p. 433.
- [38] Jonas Moćkus. “On Bayesian methods for seeking the extremum”. In: *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*. Springer. 1975, pp. 400–404.
- [39] P Mogensen and A Riseth. “Optim: A mathematical optimization package for Julia”. In: *Journal of Open Source Software* 3.24 (2018).
- [40] B Chandra Mohan and R Baskaran. “A survey: Ant Colony Optimization based recent research and implementation on several engineering domain”. In: *Expert Syst. Appl.* 39.4 (2012), pp. 4618–4627.
- [41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

-
- [42] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. “A consensus-based model for global optimization and its mean-field limit”. In: *Math. Models Methods Appl. Sci.* 27.01 (2017), pp. 183–204.
- [43] LA Rastrigin. “The convergence of the random search method in the extremal control of a many parameter system”. In: *Automaton & Remote Control* 24 (1963), pp. 1337–1342.
- [44] Konstantin Riedl. “Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law”. In: *European J. Appl. Math.* (2023), pp. 1–32. DOI: [10.1017/S0956792523000293](https://doi.org/10.1017/S0956792523000293).
- [45] Konstantin Riedl, Timo Klock, Carina Geldhauser, and Massimo Fornasier. “Gradient is All You Need?” In: *Preprint arXiv:2306.09778* (2023).
- [46] Tim Roith, Leon Bungert, and Philipp Wacker. *polarcbo*. Version 1.0.1. 2023. URL: <https://github.com/PdIPS/polarcbo>.
- [47] Igor Tikh and Konstantin Riedl. *cbo-in-python*. Version 1.0. 2022. URL: <https://github.com/Igor-Tikh/cbo-in-python>.
- [48] Xin-She Yang. “Firefly algorithms for multimodal optimization”. In: *International symposium on stochastic algorithms*. Springer. 2009, pp. 169–178.

License for [CBX].

The permission to reprint and include the material is printed on the next page(s).

Content Licensing & Open Access

JOSS is a [diamond/platinum open access](#) journal. Copyright of JOSS papers is retained by submitting authors and accepted papers are subject to a [Creative Commons Attribution 4.0 International License](#).

Any code snippets included in JOSS papers are subject to the [MIT license](#) regardless of the license of the submitted software package under review.

Any use of the JOSS logo is licensed CC BY 4.0. See the [joss/logo](#) directory in the [digital-assets](#) repository for more information about it.



arXiv.org - Non-exclusive license to distribute

The URI <http://arxiv.org/licenses/nonexclusive-distrib/1.0/> is used to record the fact that the submitter granted the following license to arXiv.org on submission of an article:

- I grant arXiv.org a perpetual, non-exclusive license to distribute this article.
- I certify that I have the right to grant this license.
- I understand that submissions cannot be completely removed once accepted.
- I understand that arXiv.org reserves the right to reclassify or reject any submission.

Revision history

2004-01-16 - License above introduced as part of arXiv submission process

2007-06-21 - This HTML page created

[Contact](#)

Paper P6

Gradient is All You Need? How Consensus-Based Optimization can be Interpreted as a Stochastic Relaxation of Gradient Descent

K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier
arXiv preprint arXiv:2306.09778 (2023)

Paper Summary of [CBO&GD]³⁹

In the paper “Gradient is All You Need?” we provide a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms by interpreting CBO as a stochastic relaxation of gradient descent.

CBO is a multi-particle derivative-free optimization method, with provable convergence guarantees, capable of globally minimizing nonconvex nonsmooth functions.

Remarkably, we observe both experimentally [CBO&GD, Figure 1] and theoretically [CBO&GD, Theorem 2] that through communication of the particles, CBO exhibits an SGD-like behavior despite solely relying on evaluations of the objective function, i.e., on zero-order information. The fundamental value of such link between CBO and SGD lies in the fact that CBO is provably globally convergent to global minimizers for ample classes of nonsmooth and nonconvex objective functions [CBO-I]. Hence, on the one side, we offer a novel explanation for the success of stochastic relaxations of gradient descent by furnishing useful and precise insights that explain how problem-tailored stochastic perturbations of gradient descent (like the ones induced by CBO) overcome energy barriers and reach deep levels of nonconvex functions. On the other side, and contrary to the conventional wisdom for which derivative-free methods ought to be inefficient or not to possess generalization abilities, our results unveil an intrinsic gradient descent nature of heuristics. This viewpoint furthermore complements previous insights into the working principles of CBO, which describe the dynamics in the mean-field limit through a nonlinear nonlocal partial differential equation that allows to alleviate complexities of the nonconvex function landscape [CBO-I; CBO-II]. Our proofs leverage a completely nonsmooth analysis, which combines a novel quantitative version of the Laplace principle (log-sum-exp trick) and the minimizing movement scheme (proximal iteration). In doing so, we furnish useful and precise insights that explain how stochastic perturbations of gradient descent overcome energy barriers and reach deep levels of nonconvex functions. We further widen the scope of applications of methods which—in one way or another, be it explicitly or implicitly—estimate and exploit gradients.

KR’s Contributions. Based on numerical experiments suggesting that CBO behaves GD-like in certain parameter settings, KR initiated discussions with TK and MF about investigating the trajectory of the consensus point of CBO. Independently and a bit earlier, TK proposed to study a CBO-related algorithm, which later turned out to be the consensus hopping scheme and, therefore, to that day, is fondly referred to as Timo’s scheme. Together with MF, KR sketched and worked out the individual proof steps to eventually rigorously connect CBO with GD via the consensus hopping scheme, which was rediscovered at that time. The technical details were made explicit by KR and carefully checked by and thoroughly discussed with MF. KR conducted the numerical experiments and wrote significant parts of the paper, which was then completed and refined together with MF.

³⁹In this section, we follow [CBO&GD, Abstract].

The following document is a reprint of

[CBO&GD] K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier. “Gradient is All You Need?” In: *arXiv preprint arXiv:2306.09778* (2023).

The permission to reprint and include the material is provided after the reprint.

Gradient is All You Need?

Konstantin Riedl^{1,2} Timo Klock³ Carina Geldhauser^{1,2} Massimo Fornasier^{1,2,4}

¹Technical University of Munich, Munich, Germany

²Munich Center for Machine Learning, Munich, Germany

³Deeptech Consulting, Oslo, Norway

⁴Munich Data Science Institute, Munich, Germany

{konstantin.riedl, carina.geldhauser, massimo.fornasier}@ma.tum.de
timo@deeptechconsulting.no

Abstract

In this paper we provide a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms by interpreting consensus-based optimization (CBO), a recently proposed multi-particle derivative-free optimization method, as a stochastic relaxation of gradient descent. Remarkably, we observe that through communication of the particles, CBO exhibits a stochastic gradient descent (SGD)-like behavior despite solely relying on evaluations of the objective function. The fundamental value of such link between CBO and SGD lies in the fact that *CBO is provably globally convergent to global minimizers for ample classes of nonsmooth and nonconvex objective functions*, hence, on the one side, offering a novel explanation for the success of stochastic relaxations of gradient descent. On the other side, contrary to the conventional wisdom for which zero-order methods ought to be inefficient or not to possess generalization abilities, our results unveil an intrinsic gradient descent nature of such heuristics. This viewpoint furthermore complements previous insights into the working principles of CBO, which describe the dynamics in the mean-field limit through a nonlinear nonlocal partial differential equation that allows to alleviate complexities of the nonconvex function landscape. Our proofs leverage a completely nonsmooth analysis, which combines a novel quantitative version of the Laplace principle (log-sum-exp trick) and the minimizing movement scheme (proximal iteration). In doing so, we furnish useful and precise insights that explain how stochastic perturbations of gradient descent overcome energy barriers and reach deep levels of nonconvex functions. Instructive numerical illustrations support the provided theoretical insights.

1 Introduction

Gradient-based learning algorithms, such as stochastic gradient descent (SGD), AdaGrad [1], RMSProp and Adam [2], just to name a few of the most known and advocated, have undoubtedly been one of the cornerstones of the astounding successes of machine learning [3–5] in the last decades. In particular, the efficient computation of gradients through backpropagation [6] and automatic differentiation [7] has allowed practitioners to leverage nowadays enormous amounts of data to train huge models [8]. Despite an ever-growing relevance of advancing our mathematical understanding concerning the behavior of gradient-based learning algorithms when employed to train neural networks, the fundamental reasons behind their empirical successes largely remain elusive [9] and defy our theoretical understanding [10]. Yet, over the last years, several studies have started shedding light on the peculiarities of neural network loss functions as well as the training dynamics of SGD and its variants, see, e.g., [10–24] and references therein.

In this work, we consider the more generic, ubiquitous problem of finding a global minimizer of a potentially nonsmooth and nonconvex objective function $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e., solving

$$x^* \in \arg \min_{x \in \mathbb{R}^d} \mathcal{E}(x). \quad (1)$$

We shall provide a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms for such general global optimization problem by interpreting a recently proposed multi-particle metaheuristic derivative-free (zero-order) optimization method, called consensus-based optimization (CBO) [25], as a stochastic relaxation of gradient descent (GD), see Theorem 1 below for the statement of our main result and Figure 1 for an illustration. The essential benefit of establishing such link between CBO and (S)GD lies in the fact that CBO is provably capable of achieving global convergence towards global minimizers for rich classes of nonsmooth and nonconvex objective functions [26–31], see Section 3 and in particular Theorem 4 for a review of [30, 31]. Hence, such up to now largely unexplored connection between mathematically explainable derivative-free optimization methods and gradient-based learning algorithms discloses, on the one side, a novel and complementary perspective on why stochastic relaxations of GD are so successful, and, conversely, but no less surprising, unveils an intrinsic GD nature of heuristics on the other.

Before elaborating on the aforementioned connection, let us introduce CBO in detail, distill its fundamental conceptual principles, and explain the mechanisms behind its functioning. Inspired by particle swarm optimization (PSO) [32], the method employs an interacting stochastic system of N particles X^1, \dots, X^N to explore the domain and to form consensus about the global minimizer x^* over time. More concretely, given a finite number of time steps K , a discrete time step size $\Delta t > 0$ and denoting the position of the i -th particle at time step $k \in \{0, \dots, K\}$ by X_k^i , this position is computed for user-specified parameters $\alpha, \lambda, \sigma > 0$ according to the iterative update rule

$$X_k^i = X_{k-1}^i - \Delta t \lambda (X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)) + \sigma D(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)) B_k^i, \quad (2)$$

where $\hat{\rho}_k^N$ denotes the empirical measure of the particles at time step k , i.e., $\hat{\rho}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}$. In the spirit of the exploration-exploitation philosophy of evolutionary computation techniques [33–35], the dynamics (2) of each particle is governed by two competing terms, one being stochastic, the other deterministic in nature. The first of the two terms on the right-hand side of (2) imposes a deterministic drift towards the so-called consensus point $x_\alpha^\mathcal{E}$, which is defined for a measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ by

$$x_\alpha^\mathcal{E}(\varrho) := \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\varrho)}} d\varrho(x), \quad \text{with} \quad \omega_\alpha^\mathcal{E}(x) := \exp(-\alpha \mathcal{E}(x)). \quad (3)$$

Notice that in the case $\varrho = \hat{\rho}_k^N$, Formula (3) is just a weighted (exploiting the particles' knowledge of their objective function values) convex combination of the positions X_k^i . To be precise, owed to the particular choice of Gibbs weights $\omega_\alpha^\mathcal{E}$, larger mass is attributed to particles with comparably low objective value, whereas only little mass is given to particles whose value is undesirably high. This facilitates the interpretation that $x_\alpha^\mathcal{E}(\hat{\rho}_k^N)$ is an approximation to $\arg \min_{i=1, \dots, N} \mathcal{E}(X_k^i)$, which improves as $\alpha \rightarrow \infty$ and which can be regarded as a proxy for the global minimizer x^* , based on the information currently available to the particles. Theoretically, this is justified by the log-sum-exp trick or the Laplace principle [36, 37]. Let us further remark that the particles communicate and exchange information amongst each other exclusively through sharing the consensus point $x_\alpha^\mathcal{E}$. The other term in (2) is a stochastic diffusion injecting randomness into the dynamics, thereby encoding its explorative nature. Given i.i.d. Gaussian random vectors B_k^i in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}$, each particle is subject to anisotropic noise, i.e., $D(\bullet) = \text{diag}(\bullet)$,¹ which favors exploration the farther a particle is away from the consensus point in a certain direction. In particular, the diffusive character of the dynamics vanishes over time as consensus is reached. The described exploration-exploitation mechanism can be seen as a multi-particle reincarnation of similar ones executed by simulated annealing [38–40] and the annealed Langevin dynamics [41]. System (2) is complemented with independent initial data x_0^i distributed according to a common probability measure $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$, i.e., $X_0^i = x_0^i \sim \rho_0$.

Hence, CBO distills fundamental principles from other popular and successful metaheuristics, in particular PSO and simulated annealing, but, let us emphasize, that it comes with two fundamental

¹ $\text{diag} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ denotes the operator mapping a vector onto a diagonal matrix with the vector as its diagonal.

advantages compared to these algorithms. Firstly, it outperforms such well-established methods in experiments over challenging benchmarks [42–44]. Secondly, and remarkably, it comes with solid and robust theoretical guarantees of global convergence to global minimizers [26–31]. For these reasons, it has to be considered a baseline for understanding heuristics.

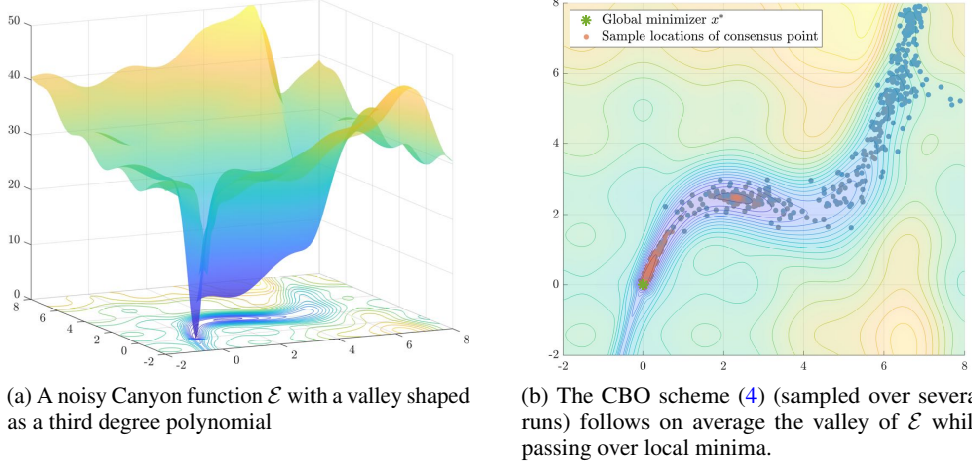


Figure 1: An illustration of the intuition that the CBO scheme (4) can be regarded as a stochastic derivative-free (zero-order) relaxation of gradient descent. To find the global minimizer x^* of the nonconvex objective function \mathcal{E} depicted in (a), we run the CBO algorithm (2) for $K = 250$ iterations with parameters $\Delta t = 0.1$, $\alpha = 100$, $\lambda = 1$ and $\sigma = 1.6$, and $N = 200$ particles, initialized i.i.d. according to $\rho_0 = \mathcal{N}((8, 8), 0.5\text{Id})$. This experiment is performed 50 times. For each run we depict in (b) the positions of the consensus points computed during the CBO algorithm (2), i.e., the iterates of the CBO scheme (4) for $k = 1, \dots, K$. The color of the individual points corresponds to time, i.e., iterates at the beginning of the scheme are plotted in blue, whereas later iterates are colored orange. We observe that, after starting close to the initial position, the trajectories of the consensus points follow the path of the valley leading to the global minimizer x^* , until it is reached. In particular, unlike gradient descent (cf. Figure 2b), the scheme (4) has the capability of jumping over locally deeper passages. Such desirable behavior is observed also for the Langevin dynamics (6) (see Figure 2c), which can be regarded as a stochastic (noisy) version of gradient descent.

An insightful theoretical understanding of the behavior of CBO is to be gained, as we are about to show, by tracing the dynamics of the consensus point $x_\alpha^\mathcal{E}$ of the CBO algorithm (2). For this purpose, let us introduce the CBO scheme as the iterates $(x_k^{\text{CBO}})_{k=0, \dots, K}$ defined according to

$$\begin{aligned} x_k^{\text{CBO}} &= x_\alpha^\mathcal{E}(\hat{\rho}_k^N), \quad \text{with} \quad \hat{\rho}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}, \\ x_0^{\text{CBO}} &= x_0 \sim \rho_0, \end{aligned} \quad (4)$$

where the particles' positions X_k^i are given by Equation (2). The main theoretical finding of this work is concerned with the observation that the iterates of the CBO scheme (4), i.e., the trajectory of the consensus point $x_\alpha^\mathcal{E}$, follow, with high probability, a stochastically perturbed GD. This is illustrated in Figure 1 below and made rigorous in the following Theorem 1, whose proof is deferred to Section 4.1.

Theorem 1 (CBO is a stochastic relaxation of GD (main result)). *Let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ be L -smooth² and satisfy minimal assumptions (summarized in Assumption 2 below). Then, for $\tau > 0$ (satisfying $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$) and with parameters $\alpha, \lambda, \sigma, \Delta t > 0$ such that $\alpha \gtrsim \frac{1}{\tau} d \log d$, the iterates $(x_k^{\text{CBO}})_{k=0, \dots, K}$ of the CBO scheme (4) follow a stochastically perturbed GD, i.e., they obey*

$$x_k^{\text{CBO}} = x_{k-1}^{\text{CBO}} - \tau \nabla \mathcal{E}(x_{k-1}^{\text{CBO}}) + g_k, \quad (5)$$

where g_k is stochastic noise fulfilling for each $k = 1, \dots, K$ with high probability the quantitative estimate $\|g_k\|_2 = \mathcal{O}(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \sqrt{\tau/\alpha} + N^{-1/2}) + \mathcal{O}(\tau)$.

²A function $f \in \mathcal{C}^1(\mathbb{R}^d)$ is L -smooth if $\|\nabla f(x) - \nabla f(x')\|_2 \leq L \|x - x'\|_2$ for all $x, x' \in \mathbb{R}^d$.

The statement of Theorem 1 has to be read with a twofold interpretation, highlighting the two sides of the same coin. First, in view of the powerful capability of CBO to converge to global minimizers for rich classes of nonsmooth and nonconvex objective functions (see Section 3 and in particular Theorem 4), Theorem 1 states that there exist stochastic relaxations of GD that are provably able to robustly and reliably overcome energy barriers and reach deep levels of nonconvex functions. Such relaxations may even be derivative-free and do not require smoothness of the objective, as is the case with CBO. Second, and conversely, against the common wisdom that derivative-free optimization heuristics search the domain mainly by random exploration and therefore ought to be inefficient, we provide evidence that such heuristics in fact work successfully in finding benign optima [45–51], precisely because they are suitable stochastic relaxations of gradient-based methods. The similar behavior of CBO and SGD is further substantiated by the following numerical illustration. While the trajectories of the CBO scheme (4) are to be seen Figure 1b, we depict for comparison in Figure 2c below the discretized dynamics of the annealed Langevin dynamics [52–54],

$$dX_t = -\nabla\mathcal{E}(X_t) dt + \sqrt{2\beta_t^{-1}} dB_t. \quad (6)$$

Both stochastic methods are capable of global minimization while overcoming energy barriers and escaping local minima. For analyses of the (annealed) Langevin dynamics we refer to [41, 55–58].

Let us now comment on a few more technical aspects of Theorem 1. First to be mentioned is that, in particular compared to Polyak-Łojasiewicz-like conditions [59] or certain families of log-Sobolev inequalities [11] required to analyze the dynamics of gradient-based methods such as (S)GD or the Langevin dynamics, the assumptions under which our statement holds are rather weak. Combined with similar assumptions being sufficient to prove global convergence of CBO (as stated in Theorem 4), this extends the class of functions, for which stochastic gradient-based methods are successful in global optimization. Secondly, the stochastic perturbations g_k in (5) are not generic as they obey precise scalings. In particular, they get tighter as soon as the discrete CBO time step size $\Delta t \ll 1$, the drift parameter $\lambda \approx 1/\Delta t \gg 1$, the noise parameter $\sigma \ll 1$, the weight parameter $\alpha \gg 1$, the number of employed particles $N \gg 1$ and the GD time step size $\tau \ll 1$. For the latter we conjecture a potential amelioration of the estimate by refining even more the quantitative Laplace principle involved in the proof of Proposition 7, which would allow to improve the order $\mathcal{O}(\tau)$ dependence of the bound for $\|g_k\|_2$. Yet, as it stands, the $\mathcal{O}(\tau)$ term is about a deterministic *bounded* perturbation of the gradient, which is possibly of smaller magnitude than the gradient. Let us stress that such bounded perturbations of gradients alone do not allow to overcome local energy barriers in general (just think of a local minimizer, around which the magnitude of gradients grows faster than the displacement: any movement from the minimizer ought necessarily to get reverted). Hence, it is the stochastic part of the perturbation that enables the convergence to global minimizers. In fact, for a moderate time step size $\Delta t > 0$, a drift parameter $\lambda > 0$ relatively small compared to $1/\Delta t$, a non-insignificant noise parameter $\sigma > 0$, a moderate value of the weight parameter $\alpha > 0$ and a modest number N of particles, CBO is factually a stochastic relaxation of GD with strong noise.

Apart from gaining primarily theoretical insights from this link, let us conclude the introduction by mentioning a further, more practical aspect of establishing such a connection. In several real-world applications, including various machine learning settings, using gradients may be undesirable or even not feasible. This can be due to the black-box nature or nonsmoothness of the objective, memory limitations constraining the use of automatic differentiation, a substantial presence of spurious local minima, or the fact that gradients carry relevant information about data, which one may wish to keep undisclosed. In machine learning, in specific, the problems of hyperparameter tuning [60, 61], convex bandits [62, 63], reinforcement learning [64], the training of sparse and pruned neural networks [65], and federated learning [66–68] stimulate interest in alternative methods to gradient-based ones. In such situations, if one still wishes to rely on a GD-like optimization behavior, Theorem 1 suggests the use of CBO (or related methods such as PSO [69]), which will be both reliable and efficient,³ with linear complexity in the number of deployed particles. We report, for instance, recent ideas in the setting of clustered federated learning [71], where CBO is leveraged to avoid reverse engineering of private data through exchange of gradients. While we do not empirically investigate the complexity of CBO or provide comparisons with the state of the art for different applications in this paper, a summary of the existing literature on this matter may be found in the footnote of Section 3.

³Needlessly to be said, but if gradients are available and inexpensive to compute, methods which exploit this information are expected to be more efficient and competitive. However, incorporating a gradient drift into CBO is possible and may bear advantages of theoretical and practical nature [70, 71].

Contributions. In view of the overwhelming empirical evidence that gradient-based learning algorithms exceed in a variety of machine learning tasks what is mathematically rigorously justified, we provide in this work a novel and surprising analytical perspective on their theoretical understanding by interpreting consensus-based optimization (CBO), which is guaranteed to globally converge to global minimizers of potentially nonsmooth and nonconvex loss functions [30, 31], as a stochastic relaxation of gradient descent (GD). Specifically, we show that in suitable scalings of its parameters, CBO — despite being a derivative-free (zero-order) optimization method — naturally approximates a stochastic gradient flow dynamics, hence implicitly behaves like a gradient-based (first-order) method, see Theorem 1 and Figure 1. To establish this connection we leverage a completely nonsmooth analysis that combines simultaneously a recently obtained quantitative version of the Laplace principle [30] (log-sum-exp trick) and the minimizing movement scheme [72] (proximal iteration [73]), which is well-known from gradient flow theory [74]. Our results furnish useful and precise insights that explain the mechanisms which enable stochastic perturbations of GD to overcome energy barriers and to reach deep levels of nonconvex objective functions, even allowing for global optimization. While the usual approach to a global analysis of (stochastic) GD requires the loss to be L -smooth and to obey the Polyak-Łojasiewicz condition, for the global convergence of CBO merely local Lipschitz continuity and a certain growth condition around the global minimizer are required [30, 31]. By establishing such surprising link between stochastic GD on the one hand and metaheuristic black-box optimization algorithms such as CBO on the other, we not just allow for complementing our theoretical understanding of successfully deployed optimization algorithms in machine learning and beyond, but we also widen the scope of applications of methods which — in one way or another, be it explicitly or implicitly — estimate and exploit gradients.

Organization. Section 2 summarizes the main assumptions under which the theoretical results of this work are valid. In Section 3 we recapitulate state-of-the-art global convergence results for CBO in the setting of potentially nonsmooth and nonconvex objective functions \mathcal{E} . Section 4 is dedicated to presenting the technical details behind the main theoretical findings of this work. We first sketch how to interpret CBO as a stochastic relaxation of GD by introducing the consensus hopping scheme, which interconnects the derivative-free with the gradient-based world in optimization. It further highlights a connection between sampling and optimization. Afterwards, the proof of our main result, Theorem 1, is provided in Section 4.1 with the central technical tools being collected in Section 4.2. The proof details together with further discussions and insights are deferred to the supplemental material. Section 5 eventually concludes the paper by discussing future perspectives. In the GitHub repository <https://github.com/> we provide the implementation of the algorithms analyzed in this work and the code used to create the visualizations.

Notation. We write $\mathcal{C}(X)$ and $\mathcal{C}^k(X)$ for the spaces of continuous and k -times continuously differentiable functions $f : X \rightarrow \mathbb{R}$, respectively. With ∇f we denote the gradient of a differentiable function f . $\mathcal{P}(\mathbb{R}^d)$, respectively $\mathcal{P}_p(\mathbb{R}^d)$, is the set containing all probability measures over \mathbb{R}^d (with finite p -th moment). $\mathcal{P}_p(\mathbb{R}^d)$ is metrized by the Wasserstein- p distance W_p , see, e.g., [75, 76]. $\mathcal{N}(m, \Sigma)$ denotes a Gaussian distribution with mean m and covariance matrix Σ .

2 Characterization of the class of objective functions

The theoretical findings of this work hold for objective functions satisfying the following conditions.

Assumption 2. *Throughout we consider objective functions $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$,*

A1 *for which there exists $x^* \in \mathbb{R}^d$ such that $\mathcal{E}(x^*) = \inf_{x \in \mathbb{R}^d} \mathcal{E}(x) =: \underline{\mathcal{E}}$,*

A2 *for which there exist $C_1, C_2 > 0$ such that*

$$|\mathcal{E}(x) - \mathcal{E}(x')| \leq C_1(\|x\|_2 + \|x'\|_2) \|x - x'\|_2 \quad \text{for all } x, x' \in \mathbb{R}^d, \quad (7)$$

$$|\mathcal{E}(x) - \underline{\mathcal{E}}| \leq C_2(1 + \|x\|_2^2) \quad \text{for all } x \in \mathbb{R}^d, \quad (8)$$

A3 *for which either $\bar{\mathcal{E}} := \sup_{x \in \mathbb{R}^d} \mathcal{E}(x) < \infty$, or for which there exist $C_3, C_4 > 0$ such that*

$$\mathcal{E}(x) - \underline{\mathcal{E}} \geq C_3 \|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^d \text{ with } \|x\|_2 \geq C_4, \quad (9)$$

A4 which are semi-convex (Λ -convex for some $\Lambda \in \mathbb{R}$), i.e., $\mathcal{E}(\bullet) - \frac{\Lambda}{2} \|\bullet\|_2^2$ is convex.

A detailed discussion of Assumptions A1–A4 may be found in Appendix B.

3 Consensus-based optimization converges globally

Let us recapitulate in this section recent global convergence results for CBO. Optimizing a nonconvex objective \mathcal{E} using the CBO dynamics (2) corresponds to an evolution of N particles in an interaction potential generated by \mathcal{E} . A global convergence analysis of this algorithm on the microscopic level proves difficult as it requires to study a system of a large number of interacting stochastic processes, which are highly correlated due to the dependence injected by communication through the consensus point $x_\alpha^\mathcal{E}$. However, with the particles being interchangeable by design of the method [25], the object of analytical interest is the empirical measure $\hat{\rho}_t^N$, whose continuous-time dynamics can be approximated, assuming propagation of chaos [77], in the mean-field limit (large-particle limit) by the solution of a nonlinear nonlocal Fokker-Planck equation of the form

$$\partial_t \rho_t = \lambda \operatorname{div}((x - x_\alpha^\mathcal{E}(\rho_t)) \rho_t) + \frac{\sigma^2}{2} \sum_{k=1}^d \partial_{kk} \left(D(x - x_\alpha^\mathcal{E}(\rho_t))_{kk}^2 \rho_t \right). \quad (10)$$

This perspective enables the use of powerful deterministic calculus tools for analysis [26]. Fornasier et al. [30, 31] recently proved that, in the mean-field limit, CBO performs a gradient descent of the Wasserstein-2 distance to a Dirac measure located at the global minimizer x^* with exponential rate. Their results are valid for large classes of optimization problems under minimal assumptions about the initialization and are in particular generic in the sense that the convergence of ρ_t is independent of the original hardness of the underlying optimization problem. More precisely it holds the following.

Theorem 3 (CBO asymptotically convexifies nonconvex problems, [31, Theorem 2]). *Fix $\varepsilon > 0$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 and $\|x - x^*\|_\infty \leq (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu / \eta$ for all $x \in \mathbb{R}^d$ with constants $\eta, \nu > 0$. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ with $x^* \in \operatorname{supp} \rho_0$. Then, for any $\gamma \in (0, 1)$ and parameters $\lambda, \sigma > 0$ with $2\lambda > \sigma^2$, there exists $\alpha_0 = \alpha_0(\varepsilon, \gamma, \lambda, \sigma, d, \nu, \eta, \rho_0)$ such that for all $\alpha \geq \alpha_0$ a weak solution $(\rho_t)_{t \in [0, T^*]}$ to (10) satisfies $\min_{t \in [0, T^*]} W_2^2(\rho_t, \delta_{x^*}) \leq \varepsilon$, where $T^* = \frac{1}{(1-\gamma)(2\lambda - \sigma^2)} \log(W_2^2(\rho_0, \delta_{x^*})/\varepsilon)$. Furthermore, until the accuracy ε is reached, it holds*

$$W_2^2(\rho_t, \delta_{x^*}) \leq W_2^2(\rho_0, \delta_{x^*}) \exp(-(1-\gamma)(2\lambda - \sigma^2)t). \quad (11)$$

While Theorem 3 captures a canonical convexification of a large class of nonconvex optimization problems as the number of optimizing particles of CBO approaches infinity, it fails to explain empirically observed successes of the method using just few particles for high-dimensional problems coming from signal processing [29, 70] and machine learning [27, 31, 29, 70, 71].⁴ However, by ensuring that propagation of chaos [77] holds, Fornasier et al. [30] quantify that the fluctuations of the empirical measure $\hat{\rho}_t^N$ around ρ_t are of order $\mathcal{O}(N^{-1/2})$ for any finite time horizon. This allows to obtain probabilistic global convergence guarantees of the CBO dynamics (2) of the following kind.

Theorem 4 (Global CBO convergence, [30, Theorem 13]). *Let $\varepsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3 and consider valid the assumptions of Theorem 3. Then, with probability larger than $1 - (\delta + \varepsilon_{\text{total}}^{-1}(C_D \Delta t + C_{\text{MF}} N^{-1} + \varepsilon))$, the final iterations $(X_K^i)_{i=1, \dots, N}$ of (2) fulfill*

$$\left\| \frac{1}{N} \sum_{i=1}^N X_K^i - x^* \right\|_2^2 \leq \varepsilon_{\text{total}}, \quad (12)$$

where, besides problem-dependent constants, $C_D = C_D(d, N, T^*, \delta^{-1})$ and $C_{\text{MF}} = C_{\text{MF}}(\alpha, T^*, \delta^{-1})$.

Despite the results of this section requiring the global minimizer x^* to be unique, there exists a polarized variant of CBO [78] capable of finding multiple global minimizers at the same time.

⁴[29] applies CBO for a phase retrieval problem, robust subspace detection, and the robust computation of eigenfaces; [70] solves a compressed sensing task; [27, 31, 70] train shallow neural networks; [71] devises FedCBO to solve clustered federated learning problems while ensuring maximal data privacy

Remark 5. A conceptually similar philosophy has been taken recently by Mei et al. [10], Rotskoff and Vanden-Eijnden [12], Chizat and Bach [11], and Sirignano and Spiliopoulos [13] to explain the generalization capabilities of over-parameterized neural networks. Leveraging that the mean-field description (w.r.t. the number of neurons) of the SGD learning dynamics is captured by a nonlinear PDE that admits a gradient flow structure on $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, they show that, remarkably, original complexities of the loss landscape are alleviated in this scaling. Together with a quantification of the fluctuations of the empirical neuron distribution around this mean-field limit, they derive convergence results for SGD for sufficiently large networks with optimal generalization error.

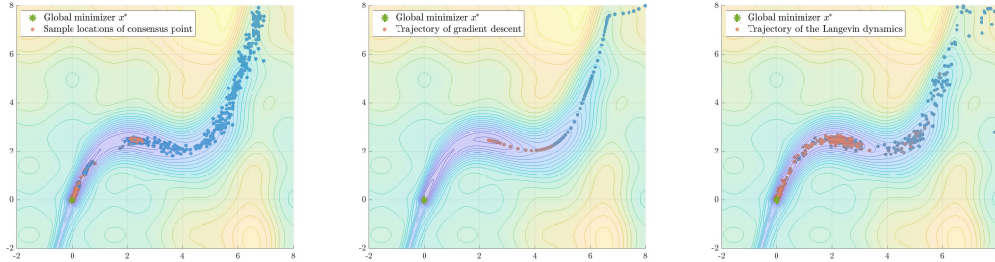
4 Consensus-based optimization is a stochastic relaxation of gradient descent

In this section we present the technical details behind the main theoretical result of this work, Theorem 1, i.e., we explain how to establish a connection between the CBO scheme (4), which captures the flow of the derivative-free CBO dynamics (2), and GD.

From CBO to consensus hopping. Let us envision for the moment the movement of the particles during the CBO dynamics (2). At every time step k , after having computed $x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)$, each particle moves a $\Delta t \lambda$ fraction of its distance towards this consensus point, before being perturbed by stochastic noise. As we let $\lambda \rightarrow 1/\Delta t$, the particles' velocities increase, until, in the case $\lambda = 1/\Delta t$, each of them hops directly to the previously computed consensus point, followed by a random fluctuation. Put differently, we are left with a numerical scheme, which, at time step k , samples N particles around the old iterate in order to subsequently compute as new iterate the consensus point (3) of the empirical measure of the samples. Such algorithm is precisely a Monte Carlo approximation of the consensus hopping (CH) scheme with iterates $(x_k^{\text{CH}})_{k=0,\dots,K}$ defined by

$$\begin{aligned} x_k^{\text{CH}} &= x_\alpha^\mathcal{E}(\mu_k), \quad \text{with} \quad \mu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id}), \\ x_0^{\text{CH}} &= x_0. \end{aligned} \tag{13}$$

Theorem 6 in Section 4.2 makes this intuition rigorous by quantifying the approximation quality between CBO and CH in terms of the parameters of the schemes. Sample trajectories of the CH scheme are depicted in Figure 2a.



(a) The CH scheme (13) (sampled over several runs) follows on average the valley of \mathcal{E} and can occasionally escape local minima.

(b) Gradient descent gets stuck in a local minimum of \mathcal{E} .

(c) The Langevin dynamics (6) (sampled over several runs) follows on average the valley of \mathcal{E} and escapes local minima.

Figure 2: An illustrative comparison between the algorithms discussed in this work. While gradient descent (obtained as an explicit Euler time discretization of $\frac{d}{dt}x(t) = -\nabla\mathcal{E}(x(t))$ with time step size $\Delta t = 0.01$ and ran for $K = 10^4$ iterations) gets stuck in a local minimum along the valley of \mathcal{E} (see (b)), the stochastic algorithms in (a) and (c) as well as Figure 1b have the capability of escaping local minima. In (a) we depict the positions of the consensus hopping scheme (13) for $K = 250$ iterations with parameters $\alpha = 100$ and $\tilde{\sigma} = 0.6$, and where we approximate the underlying measure μ_k at each step k using 200 samples. The ability of the CH scheme to escape local minima improves with larger $\tilde{\sigma}$, see Figure F.1 in Appendix F. In (c) we depict the trajectory of the overdamped Langevin dynamics (6) with $\beta_t = 0.02 \log(t+1)$ (obtained as an Euler-Maruyama time discretization of (6) with time step size $\Delta t = 0.001$ and ran for $K = 10^4$ iterations). The remaining setting is as in Figure 1, in particular, 50 individual runs of the experiment are plotted in (a) and (c).

From CH to GD. With the sampling measure μ_k assigning (in particular for small $\tilde{\sigma}$) most mass to the region close to the old iterate, the CH scheme (13) improves at every time step k its objective function value while staying near the previous iterate. A conceptually analogous behavior to such localized sampling can be achieved through penalizing the length of the step taken at time step k . This gives rise to an implicit version of the CH scheme with iterates $(\tilde{x}_k^{\text{CH}})_{k=0,\dots,K}$ given as

$$\begin{aligned} \tilde{x}_k^{\text{CH}} &= \arg \min_{x \in \mathbb{R}^d} \tilde{\mathcal{E}}_k(x), \quad \text{with} \quad \tilde{\mathcal{E}}_k(x) := \frac{1}{2\tau} \|x_{k-1}^{\text{CH}} - x\|_2^2 + \mathcal{E}(x), \\ \tilde{x}_0^{\text{CH}} &= x_0. \end{aligned} \quad (14)$$

Actually, the modulated objective $\tilde{\mathcal{E}}_k$ defined in (14) naturally appears when writing out the expression of $x_\alpha^\mathcal{E}(\mu_k)$ from (13) using that μ_k is a Gaussian. This creates a link between the sampling width $\tilde{\sigma}$ and the step size τ . The fact that the parameter τ can be seen as the step size of (14) becomes apparent when observing that the optimality condition of the k -th iterate of (14) reads $\tilde{x}_k^{\text{CH}} = x_{k-1}^{\text{CH}} - \tau \nabla \mathcal{E}(\tilde{x}_k^{\text{CH}})$, which is an implicit gradient step. Proposition 7 in Section 4.2 estimates the discrepancy between x_k^{CH} and \tilde{x}_k^{CH} employing the quantitative Laplace principle [30, Proposition 18].

Let us conclude this discussion by remarking that the scheme (14) itself is not self-consistent but requires the computation of the iterates of the CH scheme (13). For this reason we introduce the minimizing movement scheme (MMS) [72] as the iterates $(x_k^{\text{MMS}})_{k=0,\dots,K}$ given according to

$$\begin{aligned} x_k^{\text{MMS}} &= \arg \min_{x \in \mathbb{R}^d} \mathcal{E}_k(x), \quad \text{with} \quad \mathcal{E}_k(x) := \frac{1}{2\tau} \|x_{k-1}^{\text{MMS}} - x\|_2^2 + \mathcal{E}(x), \\ x_0^{\text{MMS}} &= x_0, \end{aligned} \quad (15)$$

which is known to be the discrete-time implicit Euler of the gradient flow $\frac{d}{dt}x(t) = -\nabla \mathcal{E}(x(t))$ [74].

4.1 Proof of the main result, Theorem 1

Proof of Theorem 1. From the optimality condition of the scheme $(\tilde{x}_k^{\text{CH}})_{k=1,\dots,K}$ in (14) and with the iterations $(x_k^{\text{CH}})_{k=1,\dots,K}$ as in (13), we get $(\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}) + \tau \nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) = 0$. Using this we decompose

$$x_k^{\text{CBO}} = \tilde{x}_k^{\text{CH}} + (x_k^{\text{CBO}} - \tilde{x}_k^{\text{CH}}) = x_{k-1}^{\text{CH}} - \tau \nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) + (x_k^{\text{CBO}} - \tilde{x}_k^{\text{CH}}).$$

Since $x_{k-1}^{\text{CH}} = x_{k-1}^{\text{CBO}} + (x_{k-1}^{\text{CH}} - x_{k-1}^{\text{CBO}})$ and $\nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) = \nabla \mathcal{E}(x_{k-1}^{\text{CBO}}) + (\nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) - \nabla \mathcal{E}(x_{k-1}^{\text{CBO}}))$ we can continue the former to obtain

$$x_k^{\text{CBO}} = x_{k-1}^{\text{CBO}} - \tau \nabla \mathcal{E}(x_{k-1}^{\text{CBO}}) + (x_{k-1}^{\text{CH}} - x_{k-1}^{\text{CBO}}) - \tau (\nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) - \nabla \mathcal{E}(x_{k-1}^{\text{CBO}})) + (x_k^{\text{CBO}} - \tilde{x}_k^{\text{CH}}),$$

where it remains to control the stochastic error term g_k from (5), which is comprised of the terms $g_k^1 := x_{k-1}^{\text{CH}} - x_{k-1}^{\text{CBO}}$, $g_k^2 := \tau (\nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) - \nabla \mathcal{E}(x_{k-1}^{\text{CBO}}))$ and $g_k^3 := x_k^{\text{CBO}} - \tilde{x}_k^{\text{CH}}$. By Theorem 6,

$$\|g_k^1\|_2 = \|x_{k-1}^{\text{CH}} - x_{k-1}^{\text{CBO}}\|_2 = \mathcal{O}(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \tilde{\sigma} + N^{-1/2})$$

with high probability. For g_k^2 , first notice that $\frac{1}{2\tau} \|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2^2 + \mathcal{E}(\tilde{x}_k^{\text{CH}}) \leq \mathcal{E}(x_{k-1}^{\text{CH}})$ by definition of \tilde{x}_k^{CH} , which facilitates a bound on $\|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2$ of order $\mathcal{O}(\tau)$ with high probability under A2 and by means of Remark C.7. Since \mathcal{E} is L -smooth, with the latter derivations and Theorem 6,

$$\begin{aligned} \|g_k^2\|_2 &= \tau \|\nabla \mathcal{E}(\tilde{x}_k^{\text{CH}}) - \nabla \mathcal{E}(x_{k-1}^{\text{CBO}})\|_2 \leq \tau L \|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CBO}}\|_2 \\ &\leq \tau L (\|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2 + \|x_{k-1}^{\text{CH}} - x_{k-1}^{\text{CBO}}\|_2) \\ &= \mathcal{O}(\tau^2) + \mathcal{O}(\tau(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \tilde{\sigma} + N^{-1/2})) \end{aligned}$$

with high probability. Eventually, by Theorem 6 and Proposition 7 (hence, the quantitative Laplace principle [30, Proposition 18], see Proposition E.2), it holds for a sufficiently large choice of α that

$$\begin{aligned} \|g_k^3\|_2 &= \|x_k^{\text{CBO}} - \tilde{x}_k^{\text{CH}}\|_2 \leq \|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2 + \|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2 \\ &= \mathcal{O}(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \tilde{\sigma} + N^{-1/2}) + \mathcal{O}(\tau) \end{aligned}$$

with high probability, which concludes the proof recalling that $\tilde{\sigma}^2 = \tau/(2\alpha)$ as of Proposition 7. \square

4.2 Technical details connecting CBO with GD via the CH scheme (13)

We now make rigorous what was described colloquially at the beginning of this section. The proofs of the results below are deferred to Appendices D and E. \mathcal{M} is the moment bound from Remark C.7.

CBO is a stochastic relaxation of CH. Theorem 6 explains how the CBO scheme (4) can be interpreted as a stochastic relaxation of the CH scheme (13).

Theorem 6 (CBO relaxes CH). Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. We denote by $(x_k^{\text{CBO}})_{k=0, \dots, K}$ the iterates of the CBO scheme (4) and by $(x_k^{\text{CH}})_{k=0, \dots, K}$ the ones of the CH scheme (13). Then, with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that

$$\|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \leq \varepsilon^{-1} C (|\lambda - 1/\Delta t|^2 + \sigma^2 \Delta t + \tilde{\sigma}^2 + N^{-1}) \quad (16)$$

with a constant $C = C(\delta^{-1}, \Delta t, d, \alpha, \lambda, \sigma, b_1, b_2, C_1, C_2, K, \mathcal{M})$.

CH behaves like a gradient-based method. Since by definition of the iterates \tilde{x}_k^{CH} in (14), it holds $\tilde{x}_k^{\text{CH}} = x_{k-1}^{\text{CH}} - \tau \nabla \mathcal{E}(\tilde{x}_k^{\text{CH}})$, Proposition 7 constitutes that (granted a sufficiently large choice of α and a suitably small choice of $\tilde{\sigma}$) the CH scheme (13) performs a gradient step at every time step k .

Proposition 7 (CH performs gradient steps). Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A4. We denote by $(x_k^{\text{CH}})_{k=0, \dots, K}$ the iterations of the CH scheme (13) and by $(\tilde{x}_k^{\text{CH}})_{k=0, \dots, K}$ the ones of the scheme (14). Moreover, assume that the parameters α, τ and $\tilde{\sigma}$ are such that $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$, $\alpha \gtrsim \frac{1}{\tau} d \log d$ is sufficiently large and $\tilde{\sigma}^2 = \tau/(2\alpha)$. Then, with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that

$$\|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \leq \varepsilon^{-1} c \tau^2 \quad (17)$$

with a constant $c = c(\delta^{-1}, C_1, \mathcal{M})$.

The proof of Proposition 7 is based on the quantitative Laplace principle [30, Proposition 18] (see also Proposition E.2). We conjecture that a refinement thereof may allow to control the error in (17) just through α and $\tilde{\sigma}$ without creating a dependence on τ . Nevertheless, the bound is sufficient to suggest a gradient-like behavior of the CH scheme (13) (see the discussion after Theorem 1).

Combining Proposition 7 with a stability argument for the MMS and applying Grönwall's inequality allows to control in Theorem 8 the divergence between the CH scheme (13) and the MMS (15).

Theorem 8 (CH relaxes a gradient flow). Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A4. We denote by $(x_k^{\text{CH}})_{k=0, \dots, K}$ the iterations of the CH scheme (13) and by $(x_k^{\text{MMS}})_{k=0, \dots, K}$ the ones of the MMS (15). Moreover, assume that the parameters α, τ and $\tilde{\sigma}$ are such that $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$, $\alpha \gtrsim \frac{1}{\tau} d \log d$ is sufficiently large and $\tilde{\sigma}^2 = \tau/(2\alpha)$. Then, with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that

$$\|x_k^{\text{CH}} - x_k^{\text{MMS}}\|_2^2 \leq \varepsilon^{-1} c (1 + \vartheta^{-1}) \tau^2 \sum_{\ell=0}^{k-1} \left(\frac{1 + \vartheta}{(1 + \tau\Lambda)^2} \right)^\ell \quad (18)$$

for any $\vartheta \in (0, 1)$ and with a constant $c = c(\delta^{-1}, C_1, \mathcal{M})$.

Corollary 9. Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A4 with $\Lambda > 0$. Then, in the setting of Theorem 8 and with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that

$$\|x_k^{\text{CH}} - x_k^{\text{MMS}}\|_2^2 \leq \varepsilon^{-1} c (1 + \vartheta^{-1}) \tau^2 \frac{(1 + \tau\Lambda)^2}{(1 + \tau\Lambda)^2 - (1 + \vartheta)}. \quad (19)$$

5 Conclusions

In this paper we provided a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms by showing that consensus-based optimization (CBO), an intrinsically derivative-free optimization method guaranteed to globally converge to global minimizers of potentially nonsmooth and nonconvex loss functions, implicitly behaves like a gradient-based method. This allows to interpret CBO as a stochastic relaxation of gradient descent. Besides forging such unexpected link and thereby driving forward our theoretical understanding of both gradient-based

learning methods and metaheuristic black-box optimization algorithms, we widen the scope of applications of methods which — in one way or another, be it explicitly or implicitly — estimate and exploit gradients. In particular, we believe these insights to bear the potential for designing efficient and reliable training methods which behave like first-order methods while not relying on the ability of computing gradients. Potential areas of application in machine learning may include the usage of nonsmooth losses, hyperparameter tuning, convex bandits, reinforcement learning, the training of sparse and pruned neural networks, or federated learning.

An analogous analysis approach may be carried over to second-order methods (with momentum), allowing to establish a link between Adam [2] and the well-known particle swarm optimization method [32], which is related to CBO through a zero-inertia limit [42, 69]. Together with recent observations [79] based on tools from kinetic theory that simulated annealing [38–40] is related to the Langevin dynamics [52–54], this would strengthen even further the surprising and yet largely unexplored link between gradient-based learning algorithms and derivative-free metaheuristic optimization methods. Beyond that we envisage the likely connections between consensus-based sampling [80] and log-concave sampling or sampling by Langevin flows [81–84].

Acknowledgments and Disclosure of Funding

The authors would like to profusely thank Hui Huang, Giuseppe Savaré, and Alessandro Scagliotti for many fruitful and stimulating discussions about the topic.

This work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content. KR further acknowledges the financial support from the Technical University of Munich – Institute for Ethics in Artificial Intelligence (IEAI).

References

- [1] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.
- [2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [3] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.
- [4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.
- [5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [6] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
- [7] Atılım Güneş Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.*, 18:Paper No. 153, 43, 2017.
- [8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [9] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.
- [10] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- [11] Lenaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

- [12] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.
- [13] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.
- [14] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.
- [15] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.
- [16] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.
- [17] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2603–2612. PMLR, 06–11 Aug 2017.
- [18] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4430–4438. PMLR, 2018.
- [19] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.
- [20] Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1329–1338. PMLR, 10–15 Jul 2018.
- [21] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019.
- [22] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.
- [23] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint arXiv:2011.13772*, 2020.
- [24] Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *J. Mach. Learn. Res.*, 21(136):1–48, 2020.
- [25] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.
- [26] José A. Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28(6):1037–1066, 2018.
- [27] José A. Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.*, 27(suppl.):Paper No. S5, 22, 2021.
- [28] Seung-Yeal Ha, Shi Jin, and Doheon Kim. Convergence and error estimates for time-discrete consensus-based optimization algorithms. *Numer. Math.*, 147(2):255–282, 2021.
- [29] Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.*, 22:Paper No. 237, 55, 2021.
- [30] Massimo Fornasier, Timo Klock, and Konstantin Riedl. Consensus-based optimization methods converge globally. *arXiv preprint arXiv:2103.15130*, 2021.

- [31] Massimo Fornasier, Timo Klock, and Konstantin Riedl. Convergence of anisotropic consensus-based optimization in mean-field law. In Juan Luis Jiménez Laredo, J. Ignacio Hidalgo, and Kehinde Oluwatoyin Babaagba, editors, *Applications of Evolutionary Computation*, pages 738–754, Cham, 2022. Springer International Publishing.
- [32] James Kennedy and Russel C. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.
- [33] John H. Holland. *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor, Mich., 1975.
- [34] Thomas Bäck, David B. Fogel, and Zbigniew Michalewicz, editors. *Handbook of evolutionary computation*. Institute of Physics Publishing, Bristol; Oxford University Press, New York, 1997.
- [35] David B. Fogel. *Evolutionary computation. Toward a new philosophy of machine intelligence*. IEEE Press, Piscataway, NJ, second edition, 2000.
- [36] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.
- [37] Peter D. Miller. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.
- [38] Scott Kirkpatrick, C. Daniel Gelatt Jr., and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [39] Stuart Geman and Chii-Ruey Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043, 1986.
- [40] Richard Holley and Daniel Stroock. Simulated annealing via Sobolev inequalities. *Communications in Mathematical Physics*, 115(4):553–569, 1988.
- [41] Saul B. Gelfand and Sanjoy K. Mitter. Recursive stochastic algorithms for global optimization in \mathbb{R}^d . *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.
- [42] Sara Grassi and Lorenzo Pareschi. From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 31(08):1625–1657, 2021.
- [43] Sara Grassi, Hui Huang, Lorenzo Pareschi, and Jinniao Qiu. Mean-field particle swarm optimization. *arXiv preprint arXiv:2108.00393*, 2021.
- [44] Hui Huang, Jinniao Qiu, and Konstantin Riedl. On the global convergence of particle swarm optimization methods. *Applied Mathematics & Optimization*, 88(2):30, 2023.
- [45] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.
- [46] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, 2017.
- [47] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.
- [48] Konstantinos Nikolakakis, Farzin Haddadpour, Dionysis Kalogerias, and Amin Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in Neural Information Processing Systems*, 35:31525–31541, 2022.
- [49] Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023.
- [50] Björn Engquist, Kui Ren, and Yunan Yang. Adaptive state-dependent diffusion for derivative-free optimization. *arXiv preprint arXiv:2302.04370*, 2023.
- [51] Howard Heaton, Samy Wu Fung, and Stanley Osher. Global solutions to nonconvex problems by evolution of Hamilton-Jacobi PDEs. *Communications on Applied Mathematics and Computation*, pages 1–21, 2023.

- [52] Tzoo-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in \mathbb{R}^n . *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.
- [53] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.
- [54] Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017.
- [55] David Márquez. Convergence rates for annealing diffusion processes. *The Annals of Applied Probability*, pages 1118–1139, 1997.
- [56] Mariane Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Annals of Applied Probability*, pages 10–44, 1998.
- [57] Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.
- [58] Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.
- [59] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer International Publishing, 2016.
- [60] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.
- [61] Jeremy Rapin and Olivier Teytaud. Nevergrad — a gradient-free optimization platform, 2018.
- [62] Alekh Agarwal, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24, 2011.
- [63] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, 2017.
- [64] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [65] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.*, 22(1): 10882–11005, 2021.
- [66] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.
- [67] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [68] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.
- [69] Cristina Cipriani, Hui Huang, and Jinniao Qiu. Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Appl. Math.*, 54(3):3091–3121, 2022.
- [70] Konstantín Riedl. Leveraging memory effects and gradient information in consensus-based optimization: On global convergence in mean-field law. *arXiv preprint arXiv:2211.12184*, 2022.
- [71] José A. Carrillo, Nicolas Garcia Trillos, Sixu Li, and Yuhua Zhu. FedCBO: Reaching group consensus in clustered federated learning through consensus-based optimization. *arXiv preprint arXiv:2305.02894*, 2023.
- [72] Ennio De Giorgi. New problems on minimizing movements. In *Boundary value problems for partial differential equations and applications*, volume 29 of *RMA Res. Notes Appl. Math.*, pages 81–98. Masson, Paris, 1993.
- [73] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3): 127–239, 2014.

- [74] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bull. Math. Sci.*, 7(1):87–154, 2017.
- [75] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
- [76] Cédric Villani. *Optimal transport: Old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009.
- [77] Alain-Sol Sznitman. Topics in propagation of chaos. In *École d’Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Math.*, pages 165–251. Springer, Berlin, 1991.
- [78] Leon Bungert, Philipp Wacker, and Tim Roith. Polarized consensus-based dynamics for optimization and sampling. *arXiv preprint arXiv:2211.05238*, 2022.
- [79] Lorenzo Pareschi. Boltzmann’s legacy in global optimization. Workshop on Optimal Transport, Mean-Field Models, and Machine Learning, 2023. URL <https://www.ias.tum.de/ias/event-pages/workshop-otmfml/talk-details/>.
- [80] José A. Carrillo, Franca Hoffmann, Andrew M. Stuart, and Urbain Vaes. Consensus-based sampling. *Stud. Appl. Math.*, 148(3):1069–1140, 2022.
- [81] Alan Frieze, Ravi Kannan, and Nick Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994.
- [82] Espen Bernton. Langevin Monte Carlo and JKO splitting. In *Conference on learning theory*, pages 1777–1798. PMLR, 2018.
- [83] Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.
- [84] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.
- [85] Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. Optim.*, 10(4):1116–1135, 2000.
- [86] Yi Xu, Qihang Lin, and Tianbao Yang. Adaptive SVRG methods under error bound conditions with unknown growth parameter. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’ 17*, pages 3279–3289, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [87] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.*, 165(2, Ser. A):471–507, 2017.
- [88] Ion Necoara, Yurii Nesterov, and François Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.*, 175(1-2, Ser. A):69–107, 2019.
- [89] Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2003.
- [90] Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.*, 30(14):2725–2751, 2020.
- [91] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

Supplemental Material

Supplemental material for the paper: “Gradient is All You Need?” authored by **Konstantin Riedl**, **Timo Klock**, **Carina Geldhauser**, and **Massimo Fornasier**.

This supplemental material is organized into the following six appendices.

- Appendix [A](#): Introductory facts
- Appendix [B](#): Discussion of Assumption 2
- Appendix [C](#): Boundedness of the numerical schemes
- Appendix [D](#): Proof details for Theorem 6
- Appendix [E](#): Proof details for Proposition 7 and Theorem 8
- Appendix [F](#): Additional numerical experiments

A Introductory facts

Notation. To keep the notation concise, we hide generic constants, i.e., we write $a \lesssim b$ for $a \leq cb$, if c is a constant independent of problem-dependent constants. Moreover, since we work with random variables in several instances, many equalities and inequalities hold almost surely without being mentioned explicitly. We abbreviate with i.i.d. independently and identically distributed.

We write $\|\bullet\|_2$ and $\langle \bullet, \bullet \rangle$ for the Euclidean norm and scalar product on \mathbb{R}^d , respectively. Euclidean balls are denoted by $B_r(x) := \{z \in \mathbb{R}^d : \|z - x\|_2 \leq r\}$. Moreover, we write $\|\bullet\|_\infty$ for the ℓ^∞ -norm and denote the associated ℓ^∞ -balls by $B_r^\infty(x) := \{z \in \mathbb{R}^d : \|z - x\|_\infty \leq r\}$.

For the space of continuous functions $f : X \rightarrow Y$ we write $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$ and a suitable topological space Y . For an open set $X \subset \mathbb{R}^n$ and for $Y = \mathbb{R}^m$ the space $\mathcal{C}^k(X, Y)$ contains functions $f \in \mathcal{C}(X, Y)$ that are k -times continuously differentiable. We omit Y in the real-valued case, i.e., $\mathcal{C}(X) = \mathcal{C}(X, \mathbb{R})$ and $\mathcal{C}^k(X) = \mathcal{C}^k(X, \mathbb{R})$.

The operator ∇ denotes the gradient of a function on \mathbb{R}^d .

Convex analysis. For a convex function $f \in \mathcal{C}(\mathbb{R}^d)$ the subdifferential $\partial f(x)$ at a point $x \in \mathbb{R}^d$ is the set

$$\partial f(x) = \{p \in \mathbb{R}^d : f(y) \geq f(x) + \langle p, y - x \rangle \text{ for all } y \in \mathbb{R}^d\}.$$

In the setting $f \in \mathcal{C}(\mathbb{R}^d)$, $\partial f(x)$ is closed, convex, nonempty and bounded. If $f \in \mathcal{C}^1(\mathbb{R}^d)$, $\partial f(x) = \{\nabla f(x)\}$. Moreover, it is straightforward to verify that for $x_1, x_2, p_1, p_2 \in \mathbb{R}^d$ with $p_1 \in \partial f(x_1)$ and $p_2 \in \partial f(x_2)$ it holds $\langle p_1 - p_2, x_1 - x_2 \rangle \geq 0$.

Probability measures. The set of all Borel probability measures over \mathbb{R}^d is denoted by $\mathcal{P}(\mathbb{R}^d)$. For $p > 0$, we collect measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite p -th moment $\int \|x\|_2^p d\varrho(x)$ in $\mathcal{P}_p(\mathbb{R}^d)$.

The Dirac delta δ_x for a point $x \in \mathbb{R}^d$ is a measure satisfying $\delta(B) = 1$ if $x \in B$ and $\delta(B) = 0$ if $x \notin B$ for any measurable set $B \subset \mathbb{R}^d$.

Wasserstein distance. For any $1 \leq p < \infty$, the Wasserstein- p distance between two Borel probability measures $\varrho, \varrho' \in \mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$W_p(\varrho, \varrho') = \left(\inf_{\gamma \in \Pi(\varrho, \varrho')} \int \|x - x'\|_2^p d\gamma(x, x') \right)^{1/p}, \quad (20)$$

where $\Pi(\varrho, \varrho')$ denotes the set of all couplings of (a.k.a. transport plans between) ϱ and ϱ' , i.e., the collection of all Borel probability measures over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ϱ and ϱ' on the first and second component, respectively, see, e.g., [75, 76]. $\mathcal{P}_p(\mathbb{R}^d)$ endowed with the Wasserstein- p distance W_p is a complete separable metric space [75, Proposition 7.1.5].

A generalized triangle-type inequality. It holds for $p, J \in \mathbb{N}$ by Hölder's inequality

$$\left| \sum_{j=1}^J a_j \right|^p \leq J^{p-1} \sum_{j=1}^J |a_j|^p. \quad (21)$$

A discrete variant of Grönwall's inequality. If $z_k \leq az_{k-1} + b$ with $a, b \geq 0$ for all $k \geq 1$, then

$$z_k \leq a^k z_0 + b \sum_{\ell=0}^{k-1} a^\ell \leq a^k z_0 + b \prod_{\ell=1}^{k-1} (1+a) \leq a^k z_0 + be^{a(k-1)} \quad (22)$$

for all $k \geq 1$. Notice that, while the first inequality in (22) is as sharp as the initial estimates, the remaining two inequalities are rather rough upper bounds.

B Discussion of Assumption 2

Assumption A1 requires that the continuous objective function \mathcal{E} attains its globally minimal value $\underline{\mathcal{E}}$ at some $x^* \in \mathbb{R}^d$. This does in particular not exclude objectives with multiple global minimizers.

Remark B.1. For the global convergence results [30, 31] of CBO (which we recapitulated in Section 3), however, uniqueness of the global minimizer x^* is required and implied by an additional coercivity condition of the form $\|x - x^*\|_\infty \leq (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu / \eta$, which has to hold for all $x \in \mathbb{R}^d$ with constants $\eta, \nu > 0$. It can be regarded as a tractability condition of the energy landscape of \mathcal{E} and is also known as the inverse continuity property from [29] or as the error bound condition from [85–88]. Actually, as stated in [30, Assumption A2], it is sufficient if such coercivity condition holds locally around the unique global minimizer x^* , provided that in the farfield, \mathcal{E} is well above $\underline{\mathcal{E}}$. More precisely, for the results of Section 3 to hold, it is sufficient if $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfies

$$\begin{aligned} \|x - x^*\|_\infty &\leq \frac{1}{\eta} (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu \quad \text{for all } x \in B_{R_0}^\infty(x^*) \\ \mathcal{E}(x) - \underline{\mathcal{E}} &> \mathcal{E}_\infty \quad \text{for all } x \in (B_{R_0}^\infty(x^*))^c \end{aligned}$$

with constants $\eta, \nu, \mathcal{E}_\infty, R_0 > 0$.

To deploy CBO in the setting of objective functions with several global minima, Bungert et al. [78] propose a polarized variant of CBO, which localizes the dynamics by integrating a kernel in the computation of the consensus point (3). This ensures that each particle is primarily influenced by particles close to it, allowing for the creation of clusters.

Assumptions A2 and A3 can be regarded as regularity conditions on the objective landscape of \mathcal{E} . The first part of A2, Equation (7), is a local Lipschitz condition, which ensures that the objective function does not change too quickly, assuring that the information obtained when evaluating the function is informative within a region around the point of evaluation. The second part of A2, Equation (8), controls and limits the growth of the objective in the farfield. In combination with the second option in A3, Equation (9), this forces the objective to grow quadratically in the farfield. However, note that one can always redefine the objective outside a sufficiently large ball such that both conditions are met while the other assumptions are preserved. Alternatively, the first option in A3 allows for bounded functions.

Assumption A4 requires the objective \mathcal{E} to be semi-convex with parameter $\Lambda \in \mathbb{R}$. For $\Lambda > 0$, Λ -convexity is stronger than convexity (strong convexity with parameter Λ). For $\Lambda < 0$, semi-convexity is weaker, i.e., potentially nonconvex functions \mathcal{E} are included in the definition. The class of semi-convex functions is typical in the literature of gradient flows, since their general theory extends from the convex to this more general setting [74]. One particular property, which we shall exploit in this work, is that for such functions the time discretization of a gradient flow, potentially for a small step size, defined through an iterated scheme, called minimizing movement scheme [72], is well-defined. However, while semi-convexity is useful to ensure the well-posedness of gradient flows, it is not sufficient to obtain convergence to global minimizers. Other properties such as the Polyak-Łojasiewicz condition [59] or the log-Sobolev inequalities governing the flow of the Langevin dynamics [11] may be necessary.

C Boundedness of the numerical schemes

Before showing the boundedness in expectation of the numerical schemes (4), (13), (15) and (14) over time in Sections C.1–C.4, respectively, let us first recall from [26, Lemma 3.3] an estimate on the consensus point (3), which facilitates the subsequent proofs.

Lemma C.1 (Boundedness of consensus point $x_\alpha^\mathcal{E}$). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\varrho \in \mathcal{P}_2(\mathbb{R}^d)$. Then it holds*

$$\|x_\alpha^\mathcal{E}(\varrho)\|_2^2 \leq b_1 + b_2 \int \|x\|_2^2 d\varrho(x)$$

with constants $b_1 = 0$ and $b_2 = b_2(\alpha, \mathcal{E}, \bar{\mathcal{E}}) > 0$ in case the first condition of A3 holds and with $b_i = b_i(\alpha, C_2, C_3, C_4) > 0$ for $i = 1, 2$ as given in (23) in case of the second condition of A3.

Proof. In case the first condition of A3 holds, we have by definition of the consensus point $x_\alpha^\mathcal{E}$ in (3) and Jensen's inequality

$$\|x_\alpha^\mathcal{E}(\varrho)\|_2^2 \leq \int \|x\|_2^2 \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\varrho)}} d\varrho(x) \leq e^{\alpha(\bar{\mathcal{E}}-\mathcal{E})} \int \|x\|_2^2 d\varrho(x).$$

In case of the second condition of A3, the statement follows from [26, Lemma 3.3] with constants

$$b_1 = C_4^2 + b_2 \quad \text{and} \quad b_2 = 2 \frac{C_2}{C_3} \left(1 + \frac{1}{\alpha C_3} \frac{1}{C_4^2} \right), \quad (23)$$

which concludes the proof. \square

With this estimate we have all necessary tools at hand to prove the boundedness of the numerical schemes investigated in this paper.

C.1 Boundedness of the consensus-based optimization (CBO) dynamics (2) and (4)

Let us remind the reader that the iterates $(x_k^{\text{CBO}})_{k=0, \dots, K}$ of the consensus-based optimization (CBO) scheme (4) are defined by

$$\begin{aligned} x_k^{\text{CBO}} &= x_\alpha^\mathcal{E}(\hat{\rho}_k^N), \quad \text{with} \quad \hat{\rho}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}, \\ x_0^{\text{CBO}} &= x_0 \sim \rho_0, \end{aligned}$$

where the iterates $((X_k^i)_{k=0, \dots, K})_{i=1, \dots, N}$ are given as in (2) by

$$\begin{aligned} X_k^i &= X_{k-1}^i - \Delta t \lambda (X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)) + \sigma D(X_{k-1}^i - x_\alpha^\mathcal{E}(\hat{\rho}_{k-1}^N)) B_k^i, \\ X_0^i &= x_0^i \sim \rho_0 \end{aligned}$$

with B_k^i being i.i.d. Gaussian random vectors in \mathbb{R}^d with zero mean and covariance matrix $\Delta t \text{Id}$ for $k = 0, \dots, K$ and $i = 1, \dots, N$, i.e., $B_k^i \sim \mathcal{N}(0, \Delta t \text{Id})$.

Lemma C.2 (Boundedness of the CBO dynamics (2) and the CBO scheme (4)). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the empirical random measures $(\hat{\rho}_k^N)_{k=0, \dots, K}$ and the iterates $(X_k^i)_{k=0, \dots, K}$ of (2) it holds*

$$\mathbb{E} \max_{k=0, \dots, K} \int \|x\|_2^4 d\hat{\rho}_k^N(x) \leq \mathcal{M}^{\text{CBO}} \quad \text{and} \quad \max_{i=1, \dots, N} \mathbb{E} \max_{k=0, \dots, K} \|X_k^i\|_2^4 \leq \mathcal{M}^{\text{CBO}}$$

with a constant $\mathcal{M}^{\text{CBO}} = \mathcal{M}^{\text{CBO}}(\lambda, \sigma, d, b_1, b_2, K \Delta t, K, \rho_0) > 0$. Moreover, for the iterates $(x_k^{\text{CBO}})_{k=0, \dots, K}$ of (4) it holds

$$\mathbb{E} \max_{k=0, \dots, K} \|x_k^{\text{CBO}}\|_2^4 \leq \mathcal{M}^{\text{CBO}}.$$

Proof. We first note that X_k^i as defined iteratively in (2) satisfies

$$X_k^i = X_0^i - \Delta t \lambda \sum_{\ell=1}^k (X_{\ell-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)) + \sigma \sum_{\ell=1}^k D(X_{\ell-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)) B_\ell^i$$

and that for any $k = 1, \dots, K$ by means of the standard inequality (21) for $p = 4$ and $J = 3$ we have

$$\begin{aligned} \max_{\ell=0, \dots, k} \|X_\ell^i\|_2^4 &\lesssim \|X_0^i\|_2^4 + (\Delta t \lambda)^4 \max_{\ell=1, \dots, k} \left\| \sum_{m=1}^{\ell} (X_{m-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{m-1}^N)) \right\|_2^4 \\ &\quad + \sigma^4 \max_{\ell=1, \dots, k} \left\| \sum_{m=1}^{\ell} D(X_{m-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{m-1}^N)) B_m^i \right\|_2^4. \end{aligned} \quad (24)$$

Noticing that the random process $Y_\ell^i := \sum_{m=1}^{\ell} D(X_{m-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{m-1}^N)) B_m^i$, $\ell = 0, \dots, k$ is a martingale w.r.t. the filtration $\{\mathcal{F}_\ell = \sigma(\{X_0^i\} \cup \{B_m^i, m = 1, \dots, \ell\})\}_{\ell=0}^{k-1}$ since it satisfies $\mathbb{E}[Y_\ell^i | \mathcal{F}_{\ell-1}] = Y_{\ell-1}^i$ for $\ell = 1, \dots, k$, we can apply a discrete version of the Burkholder-Davis-Gundy inequality [89, Corollary 11.2.1] yielding

$$\mathbb{E} \max_{\ell=1, \dots, k} \left\| \sum_{m=1}^{\ell} D(X_{m-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{m-1}^N)) B_m^i \right\|_2^4 \lesssim d \mathbb{E} \sum_{j=1}^d \left(\sum_{\ell=1}^k (D(X_{\ell-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)))_{jj}^2 (B_\ell^i)_j^2 \right)^2.$$

Thus, when taking the expectation on both sides of (24) and employing Jensen's inequality, we can use the latter to obtain

$$\begin{aligned} \mathbb{E} \max_{\ell=0, \dots, k} \|X_\ell^i\|_2^4 &\lesssim \mathbb{E} \|X_0^i\|_2^4 + (\Delta t \lambda)^4 K^3 \mathbb{E} \sum_{\ell=1}^k \|X_{\ell-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)\|_2^4 \\ &\quad + \sigma^4 d K \mathbb{E} \sum_{j=1}^d \sum_{\ell=1}^k (D(X_{\ell-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)))_{jj}^4 (B_\ell^i)_j^4 \\ &\lesssim \mathbb{E} \|X_0^i\|_2^4 + (\Delta t \lambda)^4 K^3 \mathbb{E} \sum_{\ell=1}^k \left(\|X_{\ell-1}^i\|_2^4 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)\|_2^4 \right) \\ &\quad + (\Delta t)^2 \sigma^4 d K \mathbb{E} \sum_{j=1}^d \sum_{\ell=1}^k \left((X_{\ell-1}^i)_j^4 + (x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N))_j^4 \right) \\ &\lesssim (1 + (\Delta t \lambda)^4 K^3 + (\Delta t \sigma^2 d)^2 K) \mathbb{E} \sum_{\ell=1}^k \left(\|X_{\ell-1}^i\|_2^4 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)\|_2^4 \right) \\ &\lesssim (1 + \lambda^4 (K \Delta t)^4 + \sigma^4 d^2 (K \Delta t)^2) \mathbb{E} \max_{\ell=1, \dots, k} \left(\|X_{\ell-1}^i\|_2^4 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N)\|_2^4 \right) \\ &\leq C \mathbb{E} \max_{\ell=1, \dots, k} \left(\|X_{\ell-1}^i\|_2^4 + b_1^2 + b_2^2 \int \|x\|_2^4 d\widehat{\rho}_{\ell-1}^N(x) \right) \end{aligned} \quad (25)$$

with a constant $C = C(\lambda, \sigma, d, K \Delta t)$. In the second step we made use of the standard inequality (21) for $p = 4$ and $J = 2$, exploited that B_ℓ^i is independent from $D(X_{\ell-1}^i - x_\alpha^\mathcal{E}(\widehat{\rho}_{\ell-1}^N))$ for any $\ell = 1, \dots, k$ and used that the fourth moment of a Gaussian random variable $B \sim \mathcal{N}(0, 1)$ is $\mathbb{E} B^4 = 3$ (e.g., by recalling that $\mathbb{E} B^4 = \frac{d^4}{dx^4} M_B(x) \Big|_{x=0}$, where M_B denotes the moment-generating function of B). Moreover, recall that $K \Delta t$ denotes the final time horizon, and note that the last step is due to Lemma C.1. Averaging (25) over i allows to bound

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E} \max_{\ell=0, \dots, k} \|X_\ell^i\|_2^4 \leq \widetilde{C} \left(1 + \frac{1}{N} \sum_{i=1}^N \mathbb{E} \max_{\ell=1, \dots, k} \|X_{\ell-1}^i\|_2^4 \right) \quad (26)$$

with a constant $\tilde{C} = \tilde{C}(\lambda, \sigma, d, b_1, b_2, K\Delta t)$. Since $\mathbb{E} \int \|x\|_2^4 d\hat{\rho}_0^N(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|x_0^i\|_2^4$, an application of the discrete variant of Grönwall's inequality (22) yields the second inequality in

$$\begin{aligned} \mathbb{E} \max_{\ell=0, \dots, k} \int \|x\|_2^4 d\hat{\rho}_\ell^N(x) &\leq \frac{1}{N} \sum_{i=1}^N \mathbb{E} \max_{\ell=0, \dots, k} \|X_\ell^i\|_2^4 \\ &\leq \tilde{C}^k \mathbb{E} \int \|x\|_2^4 d\hat{\rho}_0^N(x) + \tilde{C} e^{\tilde{C}(k-1)}, \end{aligned} \quad (27)$$

showing that the left-hand side is bounded independently of N , which gives the first bound in the first part of the statement. Making use thereof in (25) also yields the second part after another application of Grönwall's inequality. The second part of the statement follows by noting that an application of Lemma C.1 gives

$$\begin{aligned} \mathbb{E} \max_{\ell=1, \dots, k} \|x_\ell^{\text{CBO}}\|_2^4 &= \mathbb{E} \max_{\ell=1, \dots, k} \|x_\alpha^\mathcal{E}(\hat{\rho}_\ell^N)\|_2^4 \\ &\leq 2b_1^2 + 2b_2^2 \mathbb{E} \max_{\ell=1, \dots, k} \int \|x\|_2^4 d\hat{\rho}_\ell^N(x), \end{aligned}$$

where the last expression is bounded as in (27). Recalling that $x_0^{\text{CBO}} = x_0 \sim \rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and choosing the constant \mathcal{M}^{CBO} large enough for all three estimates to hold with $k = K$ concludes the proof. \square

C.2 Boundedness of the consensus hopping scheme (13)

Let us recall that the iterates $(x_k^{\text{CH}})_{k=0, \dots, K}$ of the consensus hopping (CH) scheme (13) are defined by

$$\begin{aligned} x_k^{\text{CH}} &= x_\alpha^\mathcal{E}(\mu_k), \quad \text{with} \quad \mu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id}), \\ x_0^{\text{CH}} &= x_0. \end{aligned}$$

Lemma C.3 (Boundedness of the CH scheme (13)). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the random measures $(\mu_k)_{k=1, \dots, K}$ in (13) it holds*

$$\mathbb{E} \max_{k=1, \dots, K} \int \|x\|_2^4 d\mu_k(x) \leq \mathcal{M}^{\text{CH}}$$

with a constant $\mathcal{M}^{\text{CH}} = \mathcal{M}^{\text{CH}}(\tilde{\sigma}, d, b_1, b_2, K, \rho_0) > 0$. Moreover, for the iterates $(x_k^{\text{CH}})_{k=0, \dots, K}$ of (13) it holds

$$\mathbb{E} \max_{k=0, \dots, K} \|x_k^{\text{CH}}\|_2^4 \leq \mathcal{M}^{\text{CH}}.$$

Proof. According to the definition of the scheme (13) and with the standard inequality (21) for $p = 4$ and $J = 2$, we observe that for any $k = 2, \dots, K$ it holds

$$\begin{aligned} \int \|x\|_2^4 d\mu_k(x) &= \int \|x\|_2^4 d\mathcal{N}(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id})(x) \\ &\lesssim \|x_{k-1}^{\text{CH}}\|_2^4 + \int \|x\|_2^4 d\mathcal{N}(0, \tilde{\sigma}^2 \text{Id})(x) \\ &= \|x_\alpha^\mathcal{E}(\mu_{k-1})\|_2^4 + (d^2 + 2d)\tilde{\sigma}^4 \\ &\lesssim b_1^2 + b_2^2 \int \|x\|_2^4 d\mu_{k-1}(x) + d^2\tilde{\sigma}^4, \end{aligned}$$

where for the third step we explicitly computed that for the fourth moment of a multivariate Gaussian distribution it holds $\int \|x\|_2^4 d\mathcal{N}(0, \text{Id})(x) = d^2 + 2d$. Moreover, in the final step we employed Lemma C.1 together with Jensen's inequality. Along the same lines we have $\int \|x\|_2^4 d\mu_1(x) \lesssim \|x_0\|_2^4 + d^2\tilde{\sigma}^4$. An application of the discrete variant of Grönwall's inequality (22) therefore allows to obtain

$$\int \|x\|_2^4 d\mu_k(x) \lesssim b_2^{2k} \|x_0\|_2^4 + (b_1^2 + d^2\tilde{\sigma}^4) e^{cb_2^2(k-1)}$$

with a generic constant $c > 0$. Taking the maximum over the iterations k and the expectation w.r.t. the initial condition ρ_0 gives the first part of the statement. Recalling that $x_0^{\text{CH}} = x_0 \sim \rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$, the second part follows after an application of Lemma C.1, since

$$\begin{aligned} \mathbb{E} \max_{\ell=1,\dots,k} \|x_\ell^{\text{CH}}\|_2^4 &= \mathbb{E} \max_{\ell=1,\dots,k} \|x_\alpha^\mathcal{E}(\mu_\ell)\|_2^4 \\ &\leq 2b_1^2 + 2b_2^2 \mathbb{E} \max_{\ell=1,\dots,k} \int \|x\|_2^4 d\mu_\ell(x). \end{aligned}$$

Choosing the constant \mathcal{M}^{CH} large enough for either estimate to hold with $k = K$ concludes the proof. \square

Lemma C.4. *Let $Y_k^i \sim \mu_k$ for $i = 1, \dots, N$ and let $\widehat{\mu}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{Y_k^i}$. Then, under the assumptions of Lemma C.3, for the empirical random measures $(\widehat{\mu}_k^N)_{k=1,\dots,K}$ it holds*

$$\mathbb{E} \max_{k=1,\dots,K} \int \|x\|_2^4 d\widehat{\mu}_k^N(x) \leq \widehat{\mathcal{M}}^{\text{CH}}$$

with a constant $\widehat{\mathcal{M}}^{\text{CH}} = \widehat{\mathcal{M}}^{\text{CH}}(\tilde{\sigma}, d, b_1, b_2, K, \rho_0) > 0$.

Proof. By definition of the empirical measure $\widehat{\mu}_k^N$ it holds

$$\mathbb{E} \max_{k=1,\dots,K} \int \|x\|_2^4 d\widehat{\mu}_k^N(x) = \mathbb{E} \max_{k=1,\dots,K} \frac{1}{N} \sum_{i=1}^N \|Y_k^i\|_2^4 \leq \frac{1}{N} \sum_{i=1}^N \mathbb{E} \max_{k=1,\dots,K} \|Y_k^i\|_2^4. \quad (28)$$

Since $Y_k^i \sim \mu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id})$ for any $k = 1, \dots, K$ and $i = 1, \dots, N$, we can write $Y_k^i = x_{k-1}^{\text{CH}} + \tilde{\sigma} B_{Y,k}^i$, where $B_{Y,k}^i$ is a standard Gaussian random vector, i.e., $B_{Y,k}^i \sim \mathcal{N}(0, \text{Id})$. By means of the standard inequality (21) for $p = 4$ and $J = 2$ we thus have

$$\begin{aligned} \mathbb{E} \max_{k=1,\dots,K} \|Y_k^i\|_2^4 &\lesssim \mathbb{E} \max_{k=1,\dots,K} \|x_{k-1}^{\text{CH}}\|_2^4 + \tilde{\sigma}^4 \mathbb{E} \max_{k=1,\dots,K} \|B_{Y,k}^i\|_2^4 \\ &\leq \mathcal{M}^{\text{CH}} + K\tilde{\sigma}^4(d^2 + 2d), \end{aligned} \quad (29)$$

where in the last step we employed Lemma C.3 for the first term and bounded the maximum by the sum in the second term before using again that $\mathbb{E}\|B\|_2^4 = d^2 + 2d$ for $B \sim \mathcal{N}(0, \text{Id})$. Inserting (29) into (28) yields the claim. \square

C.3 Boundedness of the minimizing movement scheme (15)

We recall that the iterates $(x_k^{\text{MMS}})_{k=0,\dots,K}$ of the minimizing movement scheme (MMS) (15) are defined by

$$\begin{aligned} x_k^{\text{MMS}} &= \arg \min_{x \in \mathbb{R}^d} \mathcal{E}_k(x), \quad \text{with} \quad \mathcal{E}_k(x) := \frac{1}{2\tau} \|x_{k-1}^{\text{MMS}} - x\|_2^2 + \mathcal{E}(x), \\ x_0^{\text{MMS}} &= x_0. \end{aligned}$$

Lemma C.5 (Boundedness of the MMS (15)). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the iterates $(x_k^{\text{MMS}})_{k=0,\dots,K}$ of (15) it holds*

$$\mathbb{E} \max_{k=0,\dots,K} \|x_k^{\text{MMS}}\|_2^4 \leq \mathcal{M}^{\text{MMS}}$$

with a constant $\mathcal{M}^{\text{MMS}} = \mathcal{M}^{\text{MMS}}(K\tau, C_2, \rho_0) > 0$.

Proof. Since x_k^{MMS} is the minimizer of \mathcal{E}_k , see (15), a comparison with the old iterate x_{k-1}^{MMS} yields

$$\frac{1}{2\tau} \|x_{k-1}^{\text{MMS}} - x_k^{\text{MMS}}\|_2^2 + \mathcal{E}(x_k^{\text{MMS}}) \leq \mathcal{E}(x_{k-1}^{\text{MMS}})$$

for any $k = 1, \dots, K$. Using the standard inequality (21) for $p = 2$ and $J = k$, this can be utilized to obtain

$$\begin{aligned}
\|x_k^{\text{MMS}}\|_2^2 &\leq 2 \|x_0^{\text{MMS}}\|_2^2 + 2K \sum_{\ell=1}^k \|x_\ell^{\text{MMS}} - x_{\ell-1}^{\text{MMS}}\|_2^2 \\
&\leq 2 \|x_0^{\text{MMS}}\|_2^2 + 4K\tau \sum_{\ell=1}^k (\mathcal{E}(x_{\ell-1}^{\text{MMS}}) - \mathcal{E}(x_\ell^{\text{MMS}})) \\
&= 2 \|x_0^{\text{MMS}}\|_2^2 + 4K\tau (\mathcal{E}(x_0^{\text{MMS}}) - \mathcal{E}(x_k^{\text{MMS}})) \\
&\leq 2 \|x_0\|_2^2 + 4K\tau (\mathcal{E}(x_0) - \underline{\mathcal{E}}) \\
&\leq 2 \|x_0\|_2^2 + 4K\tau C_2 (1 + \|x_0\|_2^2) \\
&= 2(1 + 2K\tau C_2) \|x_0\|_2^2 + 4K\tau C_2,
\end{aligned}$$

which trivially also holds for $k = 0$. Taking the square and expectation w.r.t. the initial condition ρ_0 on both sides concludes the proof. \square

C.4 Boundedness of the implicit version of the CH scheme (14)

Let us recall that the iterates $(\tilde{x}_k^{\text{CH}})_{k=0, \dots, K}$ of the scheme (14) are defined by

$$\begin{aligned}
\tilde{x}_k^{\text{CH}} &= \arg \min_{x \in \mathbb{R}^d} \tilde{\mathcal{E}}_k(x), \quad \text{with} \quad \tilde{\mathcal{E}}_k(x) := \frac{1}{2\tau} \|x_{k-1}^{\text{CH}} - x\|_2^2 + \mathcal{E}(x), \\
\tilde{x}_0^{\text{CH}} &= x_0.
\end{aligned}$$

Lemma C.6 (Boundedness of the implicit version of the CH scheme (14)). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the iterates $(\tilde{x}_k^{\text{CH}})_{k=0, \dots, K}$ of (14) it holds*

$$\mathbb{E} \max_{k=0, \dots, K} \|\tilde{x}_k^{\text{CH}}\|_2^4 \leq \tilde{\mathcal{M}}^{\text{CH}}$$

with a constant $\tilde{\mathcal{M}}^{\text{CH}} = \tilde{\mathcal{M}}^{\text{CH}}(\tau, C_2, \mathcal{M}^{\text{CH}}) > 0$.

Proof. Since \tilde{x}_k^{CH} is the minimizer of $\tilde{\mathcal{E}}_k$, see (14), a comparison with x_{k-1}^{CH} yields

$$\frac{1}{2\tau} \|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 + \mathcal{E}(\tilde{x}_k^{\text{CH}}) \leq \mathcal{E}(x_{k-1}^{\text{CH}}).$$

This can be utilized to obtain

$$\begin{aligned}
\|\tilde{x}_k^{\text{CH}}\|_2^2 &= 2 \|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2^2 + 2 \|x_{k-1}^{\text{CH}}\|_2^2 \\
&\leq 4\tau (\mathcal{E}(x_{k-1}^{\text{CH}}) - \mathcal{E}(\tilde{x}_k^{\text{CH}})) + 2 \|x_{k-1}^{\text{CH}}\|_2^2 \\
&\leq 4\tau (\mathcal{E}(x_{k-1}^{\text{CH}}) - \underline{\mathcal{E}}) + 2 \|x_{k-1}^{\text{CH}}\|_2^2 \\
&\leq 4\tau C_2 \left(1 + \|x_{k-1}^{\text{CH}}\|_2^2\right) + 2 \|x_{k-1}^{\text{CH}}\|_2^2 \\
&= 2(1 + 2\tau C_2) \|x_{k-1}^{\text{CH}}\|_2^2 + 4\tau C_2.
\end{aligned}$$

Taking the square and expectation w.r.t. the initial condition ρ_0 on both sides concludes the proof by virtue of Lemma C.3. \square

C.5 Boundedness of all numerical schemes

Remark C.7 (Boundedness of the schemes (4), (13), (14) and (15)). *To keep the notation of the main body of the paper concise, we denote by \mathcal{M} the collective moment bound*

$$\mathcal{M} = \max \left\{ \mathcal{M}^{\text{CBO}}, \tilde{\mathcal{M}}^{\text{CBO}}, \mathcal{M}^{\text{CH}}, \tilde{\mathcal{M}}^{\text{CH}}, \widehat{\mathcal{M}}^{\text{MMS}}, \tilde{\mathcal{M}}^{\text{CH}} \right\}, \quad (30)$$

where \mathcal{M}^{CBO} , \mathcal{M}^{CH} , $\tilde{\mathcal{M}}^{\text{CH}}$, $\widehat{\mathcal{M}}^{\text{MMS}}$, and $\tilde{\mathcal{M}}^{\text{CH}}$ are as defined in Lemmas C.2, C.3, C.4, C.5, and C.6, respectively. Moreover, $\tilde{\mathcal{M}}^{\text{CBO}} = \mathcal{M}^{\text{CBO}}(1/\Delta t, \sigma, d, b_1, b_2, K\Delta t, K, \rho_0)$.

D Proof details for Theorem 6

Theorem 6 is centered around the observation that, as $\lambda \rightarrow 1/\Delta t$ in the CBO dynamics (2), the CBO scheme (4) resembles an implementation of the CH scheme (13) via sampling from the underlying distribution μ_k and computing the associated weighted empirical average. Accordingly, the proof of Theorem 6 consists of three ingredients. First, a stability estimate for the CBO dynamics (2) w.r.t. the parameter λ , see Lemma D.2. Second, a quantification of the structural difference in the noise component between the CBO scheme (4) and the CH scheme (13), and third a large deviation bound to control the sampling error associated with the Monte Carlo approximation of the CH scheme (13), see Lemma D.3.

D.1 Stability of the consensus point (3) w.r.t. the underlying measure

We first recall from [26, Lemma 3.2] in a slightly modified form a stability estimate for the consensus point (3) w.r.t. the measure from which it is computed. Loosely speaking, we show that the mapping $x_\alpha^\mathcal{E} : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}^d$ is Lipschitz-continuous in the Wasserstein-2 metric.

Lemma D.1 (Stability of the consensus point $x_\alpha^\mathcal{E}$). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, let $\varrho, \varrho' \in \mathcal{P}(\mathbb{R}^d)$ be random measures and define the cutoff function (random variable)*

$$\bar{\mathcal{I}}_M^1 = \begin{cases} 1, & \text{if } \max \left\{ \int \|\cdot\|_2^4 d\varrho, \int \|\cdot\|_2^4 d\varrho' \right\} \leq M^4, \\ 0, & \text{else.} \end{cases}$$

Then it holds

$$\|x_\alpha^\mathcal{E}(\varrho) - x_\alpha^\mathcal{E}(\varrho')\|_2 \bar{\mathcal{I}}_M^1 \leq c_0 W_2(\varrho, \varrho') \bar{\mathcal{I}}_M^1$$

with a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$.

Proof. To start with, we note that under A2 and with Jensen's inequality it holds

$$\begin{aligned} \frac{e^{-\alpha\mathcal{E}} \bar{\mathcal{I}}_M^1}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\varrho)}} &= \frac{\bar{\mathcal{I}}_M^1}{\int \exp(-\alpha(\mathcal{E}(x) - \mathcal{E})) d\varrho(x)} \leq \frac{\bar{\mathcal{I}}_M^1}{\int \exp(-\alpha C_2(1 + \|x\|_2^2)) d\varrho(x)} \\ &\leq \frac{\bar{\mathcal{I}}_M^1}{\exp(-\alpha C_2(1 + \int \|x\|_2^2 d\varrho(x)))} \leq \exp(\alpha C_2(1 + M^2)) =: c_M. \end{aligned} \quad (31)$$

An analogous statement can be obtained for the measure ϱ' .

By definition of the consensus point $x_\alpha^\mathcal{E}$ in (3), it holds for any coupling $\gamma \in \Pi(\varrho, \varrho')$ between ϱ and ϱ' by Jensen's inequality

$$\begin{aligned} \|x_\alpha^\mathcal{E}(\varrho) - x_\alpha^\mathcal{E}(\varrho')\|_2 \bar{\mathcal{I}}_M^1 &\leq \iint \left\| x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\varrho)}} - x' \frac{\omega_\alpha^\mathcal{E}(x')}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\varrho')}} \right\|_2 d\gamma(x, x') \bar{\mathcal{I}}_M^1 \\ &\leq \iint (\|T_1(x, x')\|_2 + \|T_2(x, x')\|_2 + \|T_3(x, x')\|_2) d\gamma(x, x') \bar{\mathcal{I}}_M^1, \end{aligned} \quad (32)$$

where the terms T_1, T_2 and T_3 are defined implicitly and bounded as follows. For the first term T_1 we have

$$\|T_1(x, x')\|_2 \bar{\mathcal{I}}_M^1 = \|x - x'\|_2 \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \leq c_M \|x - x'\|_2 \bar{\mathcal{I}}_M^1, \quad (33)$$

where we utilized (31) in the last step. For the second term T_2 , with A2 and again (31) we obtain

$$\begin{aligned} \|T_2(x, x')\|_2 \bar{\mathcal{I}}_M^1 &= \|x'\|_2 \frac{|\omega_\alpha^\mathcal{E}(x) - \omega_\alpha^\mathcal{E}(x')|}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \\ &\leq \|x'\|_2 \frac{\alpha e^{-\alpha\mathcal{E}} C_1 (\|x\|_2 + \|x'\|_2) \|x - x'\|_2 \bar{\mathcal{I}}_M^1}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\varrho)}} \\ &\leq \alpha c_M C_1 \|x'\|_2 (\|x\|_2 + \|x'\|_2) \|x - x'\|_2 \bar{\mathcal{I}}_M^1. \end{aligned} \quad (34)$$

Eventually, for the third term T_3 it holds by following similar steps

$$\begin{aligned}
\|T_3(x, x')\|_2 \bar{\mathcal{I}}_M^1 &= \|x'\|_2 \omega_\alpha^\varepsilon(x') \frac{\left| \|\omega_\alpha^\varepsilon\|_{L_1(\varrho')} - \|\omega_\alpha^\varepsilon\|_{L_1(\varrho)} \right|}{\|\omega_\alpha^\varepsilon\|_{L_1(\varrho)} \|\omega_\alpha^\varepsilon\|_{L_1(\varrho')}} \bar{\mathcal{I}}_M^1 \\
&\leq c_M \|x'\|_2 \frac{\iint \alpha e^{-\alpha \varepsilon} C_1 (\|x\|_2 + \|x'\|_2) \|x - x'\|_2 d\pi(x, x')}{\|\omega_\alpha^\varepsilon\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \\
&\leq \alpha c_M^2 C_1 \|x'\|_2 \iint (\|x\|_2 + \|x'\|_2) \|x - x'\|_2 d\pi(x, x') \bar{\mathcal{I}}_M^1.
\end{aligned} \tag{35}$$

Collecting the estimates (33)–(35) in (32), we obtain with Cauchy-Schwarz inequality and by exploiting the definition of $\bar{\mathcal{I}}_M^1$ that

$$\|x_\alpha^\varepsilon(\varrho) - x_\alpha^\varepsilon(\varrho')\|_2 \bar{\mathcal{I}}_M^1 \leq c_M (1 + 2\alpha(1 + c_M)C_1 M^2) \sqrt{\iint \|x - x'\|_2^2 d\gamma(x, x')} \bar{\mathcal{I}}_M^1. \tag{36}$$

Squaring both sides and optimizing over all couplings $\gamma \in \Pi(\varrho, \varrho')$ concludes the proof. \square

D.2 Stability of the CBO dynamics (2) w.r.t. the parameters λ and σ

Let us now show the stability of the CBO dynamics (2) w.r.t. its parameters, in particular, the drift and noise parameters λ and σ . For this we control in Lemma D.2 below the mismatch of the iterates of the CBO dynamics (2) for different parameters, however, provided coinciding initialization and discrete Brownian motion paths.

Lemma D.2 (Stability of the CBO dynamics (2)). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_d(\mathbb{R}^d)$. We denote by $((X_k^{i,1})_{k=0,\dots,K})_{i=1,\dots,N}$ and $((X_k^{i,2})_{k=0,\dots,K})_{i=1,\dots,N}$ solutions to (2) with parameters λ_1, σ_1 and λ_2, σ_2 , respectively. Furthermore, we write $(\hat{\rho}_k^{N,1})_{k=0,\dots,K}$ and $(\hat{\rho}_k^{N,2})_{k=0,\dots,K}$ for the associated empirical measures and introduce the cutoff function (random variable)*

$$\bar{\mathcal{I}}_{M,k}^1 = \begin{cases} 1, & \text{if } \max \left\{ \int \|\bullet\|_2^4 d\hat{\rho}_k^{N,1}, \int \|\bullet\|_2^4 d\hat{\rho}_k^{N,2} \right\} \leq M^4, \\ 0, & \text{else.} \end{cases} \tag{37}$$

Then, under the assumption of coinciding initial conditions $X_0^{i,1} = X_0^{i,2}$ for all $i = 1, \dots, N$ as well as Gaussian random vectors B_k^i for all $k = 1, \dots, K$ and all $i = 1, \dots, N$, it holds

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E} \|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \leq c_1 \left(|\lambda_1 - \lambda_2|^2 + |\sigma_1 - \sigma_2|^2 \right) e^{c_2(k-1)}$$

with constants $c_1 = c_1(\Delta t, d, b_1, b_2, M) > 0$ and $c_2 = c_2(\Delta t, d, \alpha, \lambda_2, \sigma_2, C_1, C_2, M) > 0$ for all $k \geq 1$.

Proof. Let us first remark that the cutoff function $\bar{\mathcal{I}}_{M,k}^1$ defined in (37) is adapted to the natural filtration $\{\mathcal{F}_k\}_{k=0,\dots,K}$, where \mathcal{F}_k denotes the sigma algebra generated by $\{B_\ell^i, \ell = 1, \dots, k, i = 1, \dots, N\}$. Now, using the iterative update rule (2) for $X_k^{i,1}$ and $X_k^{i,2}$ with parameters λ_1, σ_1 and λ_2, σ_2 , respectively, we obtain, by employing the standard inequality (21) for

$p = 2$ and $J = 5$, for their squared norm difference the upper bound

$$\begin{aligned}
\|X_k^{i,1} - X_k^{i,2}\|_2^2 &\lesssim \|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 + (\Delta t |\lambda_1 - \lambda_2|)^2 \left(\|X_{k-1}^{i,1}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})\|_2^2 \right) \\
&\quad + (\Delta t \lambda_2)^2 \left(\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \right) \\
&\quad + |\sigma_1 - \sigma_2|^2 \left(\|X_{k-1}^{i,1}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})\|_2^2 \right) \|B_k^i\|_2^2 \\
&\quad + \sigma_2^2 \left(\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \right) \|B_k^i\|_2^2 \\
&\lesssim \left(1 + (\Delta t \lambda_2)^2 + \sigma_2^2 \|B_k^i\|_2^2 \right) \left(\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \right) \\
&\quad + \left((\Delta t |\lambda_1 - \lambda_2|)^2 + |\sigma_1 - \sigma_2|^2 \|B_k^i\|_2^2 \right) \left(\|X_{k-1}^{i,1}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})\|_2^2 \right).
\end{aligned} \tag{38}$$

Since $\bar{\mathcal{I}}_{M,k}^1$ satisfies $\bar{\mathcal{I}}_{M,k}^1 = \bar{\mathcal{I}}_{M,k}^1 \bar{\mathcal{I}}_{M,\ell}^1$ for all $\ell \leq k$ and $\bar{\mathcal{I}}_{M,k}^1 \leq 1$, we obtain from (38) that

$$\begin{aligned}
&\|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \\
&\lesssim \left(1 + (\Delta t \lambda_2)^2 + \sigma_2^2 \|B_k^i\|_2^2 \right) \left(\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \right) \bar{\mathcal{I}}_{M,k-1}^1 \\
&\quad + \left((\Delta t |\lambda_1 - \lambda_2|)^2 + |\sigma_1 - \sigma_2|^2 \|B_k^i\|_2^2 \right) \left(\|X_{k-1}^{i,1}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})\|_2^2 \right) \bar{\mathcal{I}}_{M,k-1}^1.
\end{aligned}$$

With the random variables $X_{k-1}^{i,1}$, $X_{k-1}^{i,2}$, $x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})$, $x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})$ and $\bar{\mathcal{I}}_{M,k-1}^1$ being \mathcal{F}_{k-1} -measurable, taking the expectation w.r.t. the sampling of the random vectors B_k^i , $i = 1, \dots, N$, i.e., the conditional expectation $\mathbb{E}_k = \mathbb{E}[\cdot | \mathcal{F}_{k-1}]$, yields

$$\begin{aligned}
&\mathbb{E}_k \|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \\
&\lesssim (1 + (\Delta t \lambda_2)^2 + d\Delta t \sigma_2^2) \left(\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \right) \bar{\mathcal{I}}_{M,k-1}^1 \\
&\quad + \left((\Delta t |\lambda_1 - \lambda_2|)^2 + d\Delta t |\sigma_1 - \sigma_2|^2 \right) \left(\|X_{k-1}^{i,1}\|_2^2 + \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})\|_2^2 \right) \bar{\mathcal{I}}_{M,k-1}^1,
\end{aligned}$$

where we used the fact that $\mathbb{E}_k \|B_k^i\|_2^2 = d\Delta t$. Taking now the total expectation \mathbb{E} on both sides, we have by tower property (law of total expectation)

$$\begin{aligned}
&\mathbb{E} \|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \\
&\lesssim (1 + (\Delta t \lambda_2)^2 + d\Delta t \sigma_2^2) \left(\mathbb{E} \|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 + \mathbb{E} \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 \right) \\
&\quad + \left((\Delta t |\lambda_1 - \lambda_2|)^2 + d\Delta t |\sigma_1 - \sigma_2|^2 \right) \left(\mathbb{E} \|X_{k-1}^{i,1}\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 + \mathbb{E} \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1})\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 \right).
\end{aligned} \tag{39}$$

As a consequence of the stability estimate for the consensus point, Lemma D.1, it holds for a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$ that

$$\begin{aligned}
&\mathbb{E} \|x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^\mathcal{E}(\widehat{\rho}_{k-1}^{N,2})\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 \leq c_0 \mathbb{E} W_2^2(\widehat{\rho}_{k-1}^{N,1}, \widehat{\rho}_{k-1}^{N,2}) \bar{\mathcal{I}}_{M,k-1}^1 \\
&\leq c_0 \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1,
\end{aligned}$$

where we chose $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{X_{k-1}^{i,1}} \otimes \delta_{X_{k-1}^{i,2}}$ as viable transportation plan in Definition (20) to upper bound the Wasserstein distance in the second step. Utilizing this when averaging (39) over i gives

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \mathbb{E} \|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k}^1 &\lesssim (1 + c_0) (1 + (\Delta t \lambda_2)^2 + d\Delta t \sigma_2^2) \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|X_{k-1}^{i,1} - X_{k-1}^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 \\
&\quad + \left((\Delta t |\lambda_1 - \lambda_2|)^2 + d\Delta t |\sigma_1 - \sigma_2|^2 \right) (b_1 + (1 + b_2) M^2),
\end{aligned} \tag{40}$$

where we employed Lemma C.1 together with the definition of the cutoff function $\bar{\mathcal{I}}_{M,k-1}^1$ to obtain the bound in the second line of (40). Exploiting that $X_0^{i,1} = X_0^{i,2}$ for $i = 1, \dots, N$ by assumption, we conclude the proof by an application of the discrete variant of Grönwall's inequality (22), which proves that for all $k \geq 1$ it holds

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E} \|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \leq c_1 \left((\Delta t |\lambda_1 - \lambda_2|)^2 + d \Delta t |\sigma_1 - \sigma_2|^2 \right) e^{c_2(k-1)}$$

with constants $c_1 = c_1(b_1, b_2, M) > 0$ and $c_2 = c_2(c_0, \Delta t, d, \lambda_2, \sigma_2) > 0$. \square

D.3 A large deviation bound for the consensus point (3)

For a given measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and a set of N i.i.d. random variables $Y^i \sim \varrho$ with empirical random measure $\hat{\varrho}^N = \frac{1}{N} \sum_{i=1}^N \delta_{Y^i}$, one expects that under certain regularity assumptions it holds by the law of large numbers

$$x_\alpha^\mathcal{E}(\hat{\varrho}^N) \xrightarrow{\text{a.s.}} x_\alpha^\mathcal{E}(\varrho) \quad \text{as } N \rightarrow \infty.$$

This is made rigorous in the subsequent lemma, which is based on arguments from [90, Lemma 3.1] and [30, Lemma 23].

Lemma D.3 (Large deviation bound for the consensus point $x_\alpha^\mathcal{E}$). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, for $k = 1, \dots, K$, let $\mu_k \in \mathcal{P}(\mathbb{R}^d)$ be a random measure, let $(Y_k^i)_{i=1, \dots, N}$ be N i.i.d. random variables distributed according to μ_k , denote by $\hat{\mu}_k^N$ the empirical random measure $\hat{\mu}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{Y_k^i}$ and define the cutoff function (random variable)*

$$\bar{\mathcal{I}}_{M,k}^2 = \begin{cases} 1, & \text{if } \max \left\{ \int \|\cdot\|_2^4 d\hat{\mu}_k^N, \int \|\cdot\|_2^4 d\mu_k \right\} \leq M^4, \\ 0, & \text{else.} \end{cases} \quad (41)$$

Then it holds

$$\max_{k=1, \dots, K} \mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\mu}_k^N) - x_\alpha^\mathcal{E}(\mu_k)\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq c_3 N^{-1}$$

with a constant $c_3 = c_3(\alpha, b_1, b_2, C_2, M) > 0$.

Proof. To start with, we note that under A2 and with Jensen's inequality it holds

$$\begin{aligned} \frac{e^{-\alpha \mathcal{E}} \bar{\mathcal{I}}_{M,k}^2}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} &= \frac{\bar{\mathcal{I}}_{M,k}^2}{\frac{1}{N} \sum_{j=1}^N \exp(-\alpha(\mathcal{E}(Y_k^j) - \underline{\mathcal{E}}))} \leq \frac{\bar{\mathcal{I}}_{M,k}^2}{\frac{1}{N} \sum_{j=1}^N \exp(-\alpha C_2(1 + \|Y_k^j\|_2^2))} \\ &\leq \frac{\bar{\mathcal{I}}_{M,k}^2}{\exp(-\alpha C_2(1 + \frac{1}{N} \sum_{j=1}^N \|Y_k^j\|_2^2))} \leq \exp(\alpha C_2(1 + M^2)) =: c_M. \end{aligned} \quad (42)$$

By definition of the consensus point $x_\alpha^\mathcal{E}$ in (3), it holds

$$\begin{aligned} \|x_\alpha^\mathcal{E}(\hat{\mu}_k^N) - x_\alpha^\mathcal{E}(\mu_k)\|_2 \bar{\mathcal{I}}_{M,k}^2 &= \left\| \sum_{i=1}^N Y_k^i \frac{\omega_\alpha^\mathcal{E}(Y_k^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} - \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\mu_k)}} d\mu_k(x) \right\|_2 \bar{\mathcal{I}}_{M,k}^2 \\ &\leq (\|T_1\|_2 + \|T_2\|_2) \bar{\mathcal{I}}_{M,k}^2, \end{aligned} \quad (43)$$

where the terms T_1 and T_2 are defined implicitly and bounded as follows. For the first term T_1 we have

$$\begin{aligned} \|T_1\|_2 \bar{\mathcal{I}}_{M,k}^2 &= \left\| \sum_{i=1}^N Y_k^i \frac{\omega_\alpha^\mathcal{E}(Y_k^i)}{\sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} - \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} d\mu_k(x) \right\|_2 \bar{\mathcal{I}}_{M,k}^2 \\ &= \frac{\bar{\mathcal{I}}_{M,k}^2}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} \left\| \frac{1}{N} \sum_{i=1}^N Y_k^i \omega_\alpha^\mathcal{E}(Y_k^i) - \int x \omega_\alpha^\mathcal{E}(x) d\mu_k(x) \right\|_2 \\ &\leq c_M e^{\alpha \mathcal{E}} \left\| \frac{1}{N} \sum_{i=1}^N Y_k^i \omega_\alpha^\mathcal{E}(Y_k^i) - \int x \omega_\alpha^\mathcal{E}(x) d\mu_k(x) \right\|_2, \end{aligned} \quad (44)$$

where we utilized (42) in the last step. Similarly, for the second term T_2 we have

$$\begin{aligned}
\|T_2\|_2 \bar{\mathcal{I}}_{M,k}^2 &= \left\| \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} d\mu_k(x) - \int x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\mu_k)}} d\mu_k(x) \right\|_2 \bar{\mathcal{I}}_{M,k}^2 \\
&= \frac{\bar{\mathcal{I}}_{M,k}^2}{\frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j)} \left\| x_\alpha^\mathcal{E}(\mu_k) \right\|_2 \left| \frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j) - \int \omega_\alpha^\mathcal{E}(x) d\mu_k(x) \right|_2 \\
&\leq c_M e^{\alpha \mathcal{E}} (b_1 + b_2 M) \left| \frac{1}{N} \sum_{j=1}^N \omega_\alpha^\mathcal{E}(Y_k^j) - \int \omega_\alpha^\mathcal{E}(x) d\mu_k(x) \right|_2,
\end{aligned} \tag{45}$$

where the last step involved additionally Lemma C.1. Let us now introduce the random variables

$$Z_k^i := Y_k^i \omega_\alpha^\mathcal{E}(Y_k^i) - \int x \omega_\alpha^\mathcal{E}(x) d\mu_k(x) \quad \text{and} \quad z_k^i := \omega_\alpha^\mathcal{E}(Y_k^i) - \int \omega_\alpha^\mathcal{E}(x) d\mu_k(x),$$

respectively, which have zero expectation, and are i.i.d. for $i = 1, \dots, N$. With these definitions as well as the bounds (44) and (45) we obtain

$$\begin{aligned}
\mathbb{E} \|T_1\|_2^2 \bar{\mathcal{I}}_{M,k}^2 &\leq c_M^2 e^{2\alpha \mathcal{E}} \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N Z_k^i \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 = c_M^2 e^{2\alpha \mathcal{E}} \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \sum_{j=1}^N \langle Z_k^i, Z_k^j \rangle \bar{\mathcal{I}}_{M,k}^2 \\
&= c_M^2 e^{2\alpha \mathcal{E}} \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \|Z_k^i\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq 4c_M^2 M^2 \frac{1}{N}
\end{aligned} \tag{46}$$

and, analogously,

$$\mathbb{E} \|T_2\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq c_M^2 e^{2\alpha \mathcal{E}} (b_1 + b_2 M)^2 \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \|z_k^i\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq 4c_M^2 (b_1 + b_2 M)^2 \frac{1}{N}. \tag{47}$$

The last inequalities of (46) and (47) are due to the estimates

$$\begin{aligned}
\mathbb{E} \frac{1}{N} \sum_{i=1}^N \|Z_k^i\|_2^2 \bar{\mathcal{I}}_{M,k}^2 &\leq 2\mathbb{E} \frac{1}{N} \sum_{i=1}^N \|Y_k^i \omega_\alpha^\mathcal{E}(Y_k^i)\|_2^2 \bar{\mathcal{I}}_{M,k}^2 + 2\mathbb{E} \left\| \int x \omega_\alpha^\mathcal{E}(x) d\mu_k(x) \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \\
&\leq 2e^{-2\alpha \mathcal{E}} \mathbb{E} \frac{1}{N} \sum_{i=1}^N \|Y_k^i\|_2^2 \bar{\mathcal{I}}_{M,k}^2 + 2e^{-2\alpha \mathcal{E}} \mathbb{E} \int \|x\|_2^2 d\mu_k(x) \bar{\mathcal{I}}_{M,k}^2 \\
&\leq 4e^{-2\alpha \mathcal{E}} M^2
\end{aligned}$$

and, similarly,

$$\mathbb{E} \|z_k^1\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq 4e^{-2\alpha \mathcal{E}}.$$

Combining (46) and (47) concludes the proof. \square

Remark D.4. Alternatively to the explicit computations of Lemma D.3, the stability estimate for the consensus point, Lemma D.1, would allow to obtain

$$\max_{k=1, \dots, K} \mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\mu}_k^N) - x_\alpha^\mathcal{E}(\mu_k)\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq c_0 \max_{k=1, \dots, K} \mathbb{E} W_2^2(\hat{\mu}_k^N, \mu_k) \bar{\mathcal{I}}_{M,k}^2,$$

where $\mathbb{E} W_2^2(\hat{\mu}_k^N, \mu_k)$ can be controlled by employing [91, Theorem 1]. This, however, only gives a quantitative convergence rate of order $\mathcal{O}(N^{-2/d})$, which is affected by the curse of dimensionality. The convergence rate $\mathcal{O}(N^{-1})$ obtained in Lemma D.3 matches the one to be expected from Monte Carlo sampling.

D.4 Proof of Theorem 6

We now have all necessary tools at hand to present the detailed proof of Theorem 6.

Proof of Theorem 6. We notice that for the choice $\lambda = 1/\Delta t$ the iterative update rule of the particles of the CBO dynamics (2) becomes

$$\tilde{X}_k^i = x_\alpha^\mathcal{E}(\tilde{\rho}_{k-1}^N) + \sigma D(\tilde{X}_{k-1}^i - x_\alpha^\mathcal{E}(\tilde{\rho}_{k-1}^N)) B_k^i, \quad (48)$$

where $\tilde{\rho}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{X}_k^i}$. In this case, the associated CBO scheme (4) reads

$$\begin{aligned} \tilde{x}_k^{\text{CBO}} &= x_\alpha^\mathcal{E}(\tilde{\rho}_k^N) \quad \text{with } \tilde{\rho}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{X}_k^i}, \text{ where } \tilde{X}_k^i \sim \mathcal{N}\left(\tilde{x}_{k-1}^{\text{CBO}}, \Delta t \sigma^2 D(\tilde{X}_{k-1}^i - \tilde{x}_{k-1}^{\text{CBO}})^2\right), \\ \tilde{x}_0^{\text{CBO}} &= x_0, \end{aligned} \quad (49)$$

which resembles the CH dynamics (13) with the difference in the underlying measure on which basis the consensus point (3) is computed. Let us further denote by $\hat{\rho}_k^N$ the empirical measure $\hat{\mu}_k^N = \frac{1}{N} \sum_{i=1}^N \delta_{Y_k^i}$, where $Y_k^i \sim \mu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id})$ for $i = 1, \dots, N$, i.e., $Y_k^i = x_{k-1}^{\text{CH}} + \tilde{\sigma} B_{Y,k}^i$ with $B_{Y,k}^i$ being a standard Gaussian random vector.

To obtain the probabilistic formulation of the statement, let us denote the underlying probability space over which all considered random variables get their realizations by $(\Omega, \mathcal{F}, \mathbb{P})$ and introduce the subset Ω_M of Ω of suitably bounded random variables according to

$$\Omega_M := \left\{ \omega \in \Omega : \max_{k=0, \dots, K} \max \left\{ \int \|\cdot\|_2^4 d\hat{\rho}_k^N, \int \|\cdot\|_2^4 d\tilde{\rho}_k^N, \int \|\cdot\|_2^4 d\mu_k, \int \|\cdot\|_2^4 d\hat{\mu}_k^N \right\} \leq M^4 \right\}.$$

For the associated cutoff function (random variable) we write $\mathbb{1}_{\Omega_M}$. Moreover, let us define the cutoff functions

$$\mathcal{I}_{M,k} = \begin{cases} 1, & \text{if } \max \left\{ \int \|\cdot\|_2^4 d\hat{\rho}_k^N, \int \|\cdot\|_2^4 d\tilde{\rho}_k^N, \int \|\cdot\|_2^4 d\mu_k, \int \|\cdot\|_2^4 d\hat{\mu}_k^N \right\} \leq M^4 \text{ for all } \ell \leq k, \\ 0, & \text{else,} \end{cases} \quad (50)$$

which are adapted to the natural filtration and satisfy $\mathbb{1}_{\Omega_M} \leq \mathcal{I}_{M,k}$ as well as $\mathcal{I}_{M,k} = \mathcal{I}_{M,k} \mathcal{I}_{M,\ell}$ for all $\ell \leq k$.

We can decompose the expected squared discrepancy $\mathbb{E} \|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\Omega_M}$ between the CBO scheme (4) and the CH scheme (13) as

$$\mathbb{E} \|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k} \leq 2\mathbb{E} \|x_k^{\text{CBO}} - \tilde{x}_k^{\text{CBO}}\|_2^2 \mathcal{I}_{M,k} + 2\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k}. \quad (51)$$

In what follows we individually bound the two terms on the right-hand side of (51).

First term: Let us start with the term $\mathbb{E} \|x_k^{\text{CBO}} - \tilde{x}_k^{\text{CBO}}\|_2^2 \mathcal{I}_{M,k}$, which we bound by combining the stability estimate for the consensus point, Lemma D.1, with Lemma D.2, a stability estimate for the underlying CBO dynamics (2) w.r.t. its parameters λ and σ . Denoting the auxiliary cutoff function defined in (37) in the setting $\hat{\rho}_k^{N,1} = \hat{\rho}_k^N$ and $\hat{\rho}_k^{N,2} = \tilde{\rho}_k^N$ by $\bar{\mathcal{I}}_{M,k}^1$, we have due to Lemma D.1 the estimate

$$\begin{aligned} \mathbb{E} \|x_k^{\text{CBO}} - \tilde{x}_k^{\text{CBO}}\|_2^2 \mathcal{I}_{M,k} &= \mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\rho}_k^N) - x_\alpha^\mathcal{E}(\tilde{\rho}_k^N)\|_2^2 \mathcal{I}_{M,k} \\ &\leq \mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\rho}_k^N) - x_\alpha^\mathcal{E}(\tilde{\rho}_k^N)\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \leq c_0 \mathbb{E} W_2^2(\hat{\rho}_k^N, \tilde{\rho}_k^N) \bar{\mathcal{I}}_{M,k}^1 \end{aligned} \quad (52)$$

with a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$. In the first inequality of (52) we exploited $\mathcal{I}_{M,k} \leq \bar{\mathcal{I}}_{M,k}^1$. The Wasserstein distance appearing on the right-hand side of (52) can be upper bounded by choosing $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i} \otimes \delta_{\tilde{X}_k^i}$ as viable transportation plan in Definition (20). This constitutes the first inequality in the estimate

$$\begin{aligned} \mathbb{E} W_2^2(\hat{\rho}_k^N, \tilde{\rho}_k^N) \bar{\mathcal{I}}_{M,k}^1 &\leq \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|X_k^i - \tilde{X}_k^i\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \\ &\leq c_1 \left(|\lambda_1 - \lambda_2|^2 + |\sigma_1 - \sigma_2|^2 \right) e^{c_2(k-1)} \leq c_1 \left| \lambda - \frac{1}{\Delta t} \right|^2 e^{c_2(k-1)}, \end{aligned} \quad (53)$$

whereas the second step is a consequence of Lemma D.2 applied with $\lambda_1 = \lambda, \sigma_1 = \sigma$ and $\lambda_2 = 1/\Delta t, \sigma_2 = \sigma$ as exploited in the third step. Hence, the constants are $c_1 = c_1(\Delta t, d, b_1, b_2, M) > 0$ and $c_2 = c_2(\Delta t, d, \alpha, \lambda, \sigma, C_1, C_2, M) > 0$.

Second term: To control the term $\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k}$ we start by decomposing it according to

$$\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k} \leq 2\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_\alpha^\mathcal{E}(\hat{\mu}_k^N)\|_2^2 \mathcal{I}_{M,k} + 2\mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\mu}_k^N) - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k}, \quad (54)$$

where $\hat{\mu}_k^N$ is as introduced at the beginning of the proof. For the first summand in (54) the stability estimate for the consensus point, Lemma D.1, gives

$$\begin{aligned} \mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_\alpha^\mathcal{E}(\hat{\mu}_k^N)\|_2^2 \mathcal{I}_{M,k} &= \mathbb{E} \|x_\alpha^\mathcal{E}(\tilde{\rho}_k^N) - x_\alpha^\mathcal{E}(\hat{\mu}_k^N)\|_2^2 \mathcal{I}_{M,k} \\ &\leq c_0 \mathbb{E} W_2^2(\tilde{\rho}_k^N, \hat{\mu}_k^N) \mathcal{I}_{M,k} \end{aligned} \quad (55)$$

with a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$. By choosing $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{X}_k^i} \otimes \delta_{Y_k^i}$ as viable transportation plan in Definition (20), we can further bound

$$\mathbb{E} W_2^2(\tilde{\rho}_k^N, \hat{\mu}_k^N) \mathcal{I}_{M,k} \leq \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|\tilde{X}_k^i - Y_k^i\|_2^2 \mathcal{I}_{M,k} \quad (56)$$

and since $\tilde{X}_k^i \sim \mathcal{N}(\tilde{x}_{k-1}^{\text{CBO}}, \Delta t \sigma^2 D(\tilde{X}_{k-1}^i - \tilde{x}_{k-1}^{\text{CBO}})^2)$ and $Y_k^i \sim \mathcal{N}(x_{k-1}^{\text{CH}}, \tilde{\sigma}^2 \text{Id})$ we have

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|\tilde{X}_k^i - Y_k^i\|_2^2 \mathcal{I}_{M,k} &\leq 2\mathbb{E} \|\tilde{x}_{k-1}^{\text{CBO}} - x_{k-1}^{\text{CH}}\|_2^2 \mathcal{I}_{M,k-1} \\ &\quad + \frac{4}{N} \sum_{i=1}^N \left(\sigma^2 \mathbb{E} \|D(\tilde{X}_{k-1}^i - \tilde{x}_{k-1}^{\text{CBO}}) B_k^i\|_2^2 \mathcal{I}_{M,k-1} + \tilde{\sigma}^2 \mathbb{E} \|B_{Y,k}^i\|_2^2 \right) \\ &\leq 2\mathbb{E} \|\tilde{x}_{k-1}^{\text{CBO}} - x_{k-1}^{\text{CH}}\|_2^2 \mathcal{I}_{M,k-1} + 8\sigma^2 \Delta t (b_1 + (1+b_2)M^2) + 4\tilde{\sigma}^2. \end{aligned} \quad (57)$$

Note that in the last step we exploited the definition of the cutoff function $\mathcal{I}_{M,k}$, which allowed to derive the bound

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \mathbb{E} \|D(\tilde{X}_{k-1}^i - \tilde{x}_{k-1}^{\text{CBO}}) B_k^i\|_2^2 \mathcal{I}_{M,k-1} &\leq \frac{2}{N} \sum_{i=1}^N \mathbb{E} \left(\|\tilde{X}_{k-1}^i\|_2^2 + \|\tilde{x}_{k-1}^{\text{CBO}}\|_2^2 \right) \|B_k^i\|_2^2 \mathcal{I}_{M,k-1} \\ &\leq 2\mathbb{E} \|\tilde{x}_{k-1}^{\text{CBO}}\|_2^2 \mathcal{I}_{M,k-1} + \frac{2}{N} \sum_{i=1}^N \mathbb{E} \|\tilde{X}_{k-1}^i\|_2^2 \mathcal{I}_{M,k-1} \\ &\leq 2(b_1 + (1+b_2)M^2) \end{aligned}$$

by using Lemma C.1 and the fact that $B_k^i \sim \mathcal{N}(0, \Delta t \text{Id})$ is independent from \tilde{X}_{k-1}^i and $\tilde{x}_{k-1}^{\text{CBO}}$. Inserting (57) into (56) and this into (55) afterwards, we are left with

$$\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_\alpha^\mathcal{E}(\hat{\mu}_k^N)\|_2^2 \mathcal{I}_{M,k} \leq c \left(\mathbb{E} \|\tilde{x}_{k-1}^{\text{CBO}} - x_{k-1}^{\text{CH}}\|_2^2 \mathcal{I}_{M,k-1} + \sigma^2 \Delta t + \tilde{\sigma}^2 \right) \quad (58)$$

with a constant $c = c(c_0, b_1, b_2, M) > 0$. For the second summand in (54) we have by Lemma D.3

$$\begin{aligned} \mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\mu}_k^N) - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k} &\leq \mathbb{E} \|x_\alpha^\mathcal{E}(\hat{\mu}_k^N) - x_\alpha^\mathcal{E}(\mu_k)\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \\ &\leq c_3 N^{-1}, \end{aligned} \quad (59)$$

with $c_3 = c_3(\alpha, b_1, b_2, C_2, M) > 0$ and where $\bar{\mathcal{I}}_{M,k}^2$ is an auxiliary cutoff function as defined in (41). Combining (58) with (59) we arrive for (54) at

$$\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k} \leq c \mathbb{E} \|\tilde{x}_{k-1}^{\text{CBO}} - x_{k-1}^{\text{CH}}\|_2^2 \mathcal{I}_{M,k-1} + c\sigma^2 \Delta t + c\tilde{\sigma}^2 + c_3 N^{-1}. \quad (60)$$

An application of the discrete variant of Grönwall's inequality (22) shows that

$$\mathbb{E} \|\tilde{x}_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathcal{I}_{M,k} \leq c^k \mathbb{E} \|\tilde{x}_0^{\text{CBO}} - x_0^{\text{CH}}\|_2^2 + (c\sigma^2 \Delta t + c\tilde{\sigma}^2 + c_3 N^{-1}) e^{c(k-1)}, \quad (61)$$

where the first term vanishes as both schemes are initialized with x_0 .

Concluding step: Collecting the estimates (52) combined with (53), and (61) yields for (51) the bound

$$\begin{aligned} \mathbb{E} \|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\Omega_M} &\lesssim c_0 c_1 \left| \lambda - \frac{1}{\Delta t} \right|^2 e^{c_2(k-1)} + (c\sigma^2 \Delta t + c\tilde{\sigma}^2 + c_3 N^{-1}) e^{c(k-1)} \\ &\leq C \left(\left| \lambda - \frac{1}{\Delta t} \right|^2 + \sigma^2 \Delta t + \tilde{\sigma}^2 + c_3 N^{-1} \right), \end{aligned} \quad (62)$$

with a constant $C = C(\Delta t, d, \alpha, \lambda, \sigma, b_1, b_2, C_1, C_2, K, M) > 0$. Observe that we additionally used $\mathbb{1}_{\Omega_M} \leq \mathcal{I}_{M,k}$ as observed at the beginning.

Probabilistic formulation: We first note that with Markov's inequality we have the estimate

$$\begin{aligned} \mathbb{P}(\Omega_M^c) &= \mathbb{P} \left(\max_{k=0, \dots, K} \max \left\{ \int \|\bullet\|_2^4 d\hat{\rho}_k^N, \int \|\bullet\|_2^4 d\tilde{\rho}_k^N, \int \|\bullet\|_2^4 d\mu_k, \int \|\bullet\|_2^4 d\hat{\mu}_k^N \right\} > M^4 \right) \\ &\leq \frac{1}{M^4} \left(\mathbb{E} \max_{k=0, \dots, K} \int \|\bullet\|_2^4 d\hat{\rho}_k^N + \mathbb{E} \max_{k=0, \dots, K} \int \|\bullet\|_2^4 d\tilde{\rho}_k^N \right. \\ &\quad \left. + \mathbb{E} \max_{k=0, \dots, K} \int \|\bullet\|_2^4 d\mu_k + \mathbb{E} \max_{k=0, \dots, K} \int \|\bullet\|_2^4 d\hat{\mu}_k^N \right) \\ &\leq \frac{1}{M^4} (\mathcal{M}^{\text{CBO}} + \tilde{\mathcal{M}}^{\text{CBO}} + \mathcal{M}^{\text{CH}} + \tilde{\mathcal{M}}^{\text{CH}}), \end{aligned}$$

where the last inequality is due to Lemmas C.2, C.3 and C.4. Here, $\tilde{\mathcal{M}}^{\text{CBO}}$ represents the constant \mathcal{M}^{CBO} from Lemma C.2 in the setting where $\lambda = 1/\Delta t$, i.e., $\tilde{\mathcal{M}}^{\text{CBO}} = \mathcal{M}^{\text{CBO}}(1/\Delta t, \sigma, d, b_1, b_2, K\Delta t, K, \rho_0)$. Thus, for any $\delta \in (0, 1/2)$, a sufficiently large choice $M = M(\delta^{-1}, \mathcal{M}^{\text{CBO}}, \tilde{\mathcal{M}}^{\text{CBO}}, \mathcal{M}^{\text{CH}}, \tilde{\mathcal{M}}^{\text{CH}})$ allows to ensure $\mathbb{P}(\Omega_M^c) \leq \delta$. To conclude the proof, let us denote by $K_\varepsilon \subset \Omega$ the set, where (16) does not hold and abbreviate

$$\varepsilon = \varepsilon^{-1} C \left(\left| \lambda - \frac{1}{\Delta t} \right|^2 + \sigma^2 \Delta t + \tilde{\sigma}^2 + c_3 N^{-1} \right).$$

For the probability of this set we can estimate

$$\begin{aligned} \mathbb{P}(K_\varepsilon) &= \mathbb{P}(K_\varepsilon \cap \Omega_M) + \mathbb{P}(K_\varepsilon \cap \Omega_M^c) \leq \mathbb{P}(K_\varepsilon | \Omega_M) \mathbb{P}(\Omega_M) + \mathbb{P}(\Omega_M^c) \\ &\leq \mathbb{P}(K_\varepsilon | \Omega_M) + \delta \leq \varepsilon^{-1} \mathbb{E} \left[\|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mid \Omega_M \right] + \delta, \end{aligned} \quad (63)$$

where the last step is due to Markov's inequality. By definition of the conditional expectation we further have

$$\mathbb{E} \left[\|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mid \Omega_M \right] \leq \frac{1}{\mathbb{P}(\Omega_M)} \mathbb{E} \|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\Omega_M} \leq 2 \mathbb{E} \|x_k^{\text{CBO}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\Omega_M}.$$

Inserting now the expression from (62) concludes the proof. \square

E Proof details for Proposition 7 and Theorem 8

Proposition 7 and Theorem 8 are centered around the observation that the CH scheme (13) behaves gradient-like. To establish this, Proposition 7 exploits, by using the quantitative nonasymptotic Laplace principle (see Section E.1 and in particular Proposition E.2 for a review of [30, Proposition 18]), that one step of the implicit CH scheme (14) can be recast into the computation of a consensus point $x_\alpha^{\tilde{\mathcal{E}}}$ for an objective function of the form $\tilde{\mathcal{E}}(x) = \frac{1}{2\tau} \|\bullet - x\|_2^2 + \mathcal{E}(x)$. To prove Theorem 8, this is combined with a stability argument for the MMS (15), which relies on the Λ -convexity of \mathcal{E} (Assumption A4).

E.1 A quantitative nonasymptotic Laplace principle

The Laplace principle [36, 37] asserts that for any absolutely continuous probability measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ it holds

$$\lim_{\alpha \rightarrow \infty} \left(-\frac{1}{\alpha} \log \left(\int \exp(-\alpha \tilde{\mathcal{E}}(x)) d\varrho(x) \right) \right) = \inf_{x \in \text{supp}(\varrho)} \tilde{\mathcal{E}}(x).$$

This suggests that, as $\alpha \rightarrow \infty$, the Gibbs measure $\eta_\alpha^{\tilde{\mathcal{E}}} = \omega_\alpha^{\tilde{\mathcal{E}}}\varrho / \|\omega_\alpha^{\tilde{\mathcal{E}}}\|_{L_1(\varrho)}$ converges to a discrete probability distribution (i.e., a convex combination of Dirac measures) supported on the set of global minimizers of $\tilde{\mathcal{E}}$. However, even in the case that such minimizer is unique, it does not permit to quantify the proximity of $x_\alpha^{\tilde{\mathcal{E}}}(\varrho) = \int x d\eta_\alpha^{\tilde{\mathcal{E}}}$ (see also Equation (3)) to the minimizer of $\tilde{\mathcal{E}}$ without the following assumption (see also Remark B.1).

Definition E.1 (Inverse continuity property). *A function $\tilde{\mathcal{E}} \in C(\mathbb{R}^d)$ satisfies the ℓ^2 -inverse continuity property globally if there exist constants $\eta, \nu > 0$ such that*

$$\|x - \tilde{x}^*\|_2 \leq \frac{1}{\eta} (\tilde{\mathcal{E}}(x) - \underline{\tilde{\mathcal{E}}})^\nu \quad \text{for all } x \in \mathbb{R}^d, \quad (64)$$

where $\tilde{x}^* \in \mathbb{R}^d$ denotes the unique global minimizer of $\tilde{\mathcal{E}}$ with objective value $\underline{\tilde{\mathcal{E}}} := \inf_{x \in \mathbb{R}^d} \tilde{\mathcal{E}}(x)$.

As elaborated on in Remark B.1 for the (ℓ^∞) -inverse continuity property, it is usually sufficient if (64) holds locally around the global minimizer \tilde{x}^* . In the following Proposition E.2, however, we recall the quantitative Laplace principle in the slightly more specific form, where the ℓ^2 -inverse continuity property holds globally as required by Definition E.1. For the general version, namely in the case of functions which satisfy (64) only on an ℓ^2 -ball around \tilde{x}^* (see [30, Definition 8 (A2)] for the details), we refer to [30, Proposition 18].

Proposition E.2 (Quantitative Laplace principle). *Let $\tilde{\mathcal{E}} \in C(\mathbb{R}^d)$ satisfy the ℓ^2 -inverse continuity property in form of Definition E.1. Moreover, let $\varrho \in \mathcal{P}(\mathbb{R}^d)$. For any $r > 0$ define $\tilde{\mathcal{E}}_r := \sup_{x \in B_r(\tilde{x}^*)} \tilde{\mathcal{E}}(x) - \underline{\tilde{\mathcal{E}}}$. Then, for fixed $\alpha > 0$ it holds for any $r, q > 0$ that*

$$\|x_\alpha^{\tilde{\mathcal{E}}}(\varrho) - \tilde{x}^*\|_2 \leq \frac{(q + \tilde{\mathcal{E}}_r)^\nu}{\eta} + \frac{\exp(-\alpha q)}{\varrho(B_r(\tilde{x}^*))} \int \|x - \tilde{x}^*\|_2 d\varrho(x). \quad (65)$$

Proof. W.l.o.g. we may assume $\underline{\tilde{\mathcal{E}}} = 0$ since a constant offset to $\tilde{\mathcal{E}}$ neither affects the definition of the consensus point in (3) nor the quantities appearing on the right-hand side of (65).

By Markov's inequality it holds $\|\exp(-\alpha \tilde{\mathcal{E}})\|_{L_1(\varrho)} \geq a \varrho(\{x \in \mathbb{R}^d : \exp(-\alpha \tilde{\mathcal{E}}(x)) \geq a\})$ for any $a > 0$. With the choice $a = \exp(-\alpha \tilde{\mathcal{E}}_r)$ and noting that

$$\varrho \left(\left\{ x \in \mathbb{R}^d : \exp(-\alpha \tilde{\mathcal{E}}(x)) \geq \exp(-\alpha \tilde{\mathcal{E}}_r) \right\} \right) = \varrho \left(\left\{ x \in \mathbb{R}^d : \tilde{\mathcal{E}}(x) \leq \tilde{\mathcal{E}}_r \right\} \right) \geq \varrho(B_r(\tilde{x}^*)),$$

we obtain $\|\exp(-\alpha \tilde{\mathcal{E}})\|_{L_1(\varrho)} \geq \exp(-\alpha \tilde{\mathcal{E}}_r) \varrho(B_r(\tilde{x}^*))$. Now let $\tilde{r} \geq r > 0$. With the definition of the consensus point in (3) and by Jensen's inequality we can decompose

$$\begin{aligned} \|x_\alpha^{\tilde{\mathcal{E}}}(\varrho) - \tilde{x}^*\|_2 &\leq \int_{B_{\tilde{r}}(\tilde{x}^*)} \|x - \tilde{x}^*\|_2 \frac{\exp(-\alpha \tilde{\mathcal{E}}(x))}{\|\exp(-\alpha \tilde{\mathcal{E}})\|_{L_1(\varrho)}} d\varrho(x) \\ &\quad + \int_{(B_{\tilde{r}}(\tilde{x}^*))^c} \|x - \tilde{x}^*\|_2 \frac{\exp(-\alpha \tilde{\mathcal{E}}(x))}{\|\exp(-\alpha \tilde{\mathcal{E}})\|_{L_1(\varrho)}} d\varrho(x). \end{aligned}$$

The first term is bounded by \tilde{r} since $\|x - \tilde{x}^*\|_2 \leq \tilde{r}$ for all $x \in B_{\tilde{r}}(\tilde{x}^*)$. For the second term we use the formerly derived $\|\exp(-\alpha\tilde{\mathcal{E}})\|_{L_1(\varrho)} \geq \exp(-\alpha\tilde{\mathcal{E}}_r)\varrho(B_r(\tilde{x}^*))$ to get

$$\begin{aligned} & \int_{(B_{\tilde{r}}(\tilde{x}^*))^c} \|x - \tilde{x}^*\|_2 \frac{\exp(-\alpha\tilde{\mathcal{E}}(x))}{\|\exp(-\alpha\tilde{\mathcal{E}})\|_{L_1(\varrho)}} d\varrho(x) \\ & \leq \frac{1}{\exp(-\alpha\tilde{\mathcal{E}}_r)\varrho(B_r(\tilde{x}^*))} \int_{(B_{\tilde{r}}(\tilde{x}^*))^c} \|x - \tilde{x}^*\|_2 \exp(-\alpha\tilde{\mathcal{E}}(x)) d\varrho(x) \\ & \leq \frac{\exp\left(-\alpha\left(\inf_{x \in (B_{\tilde{r}}(\tilde{x}^*))^c} \tilde{\mathcal{E}}(x) - \tilde{\mathcal{E}}_r\right)\right)}{\varrho(B_r(\tilde{x}^*))} \int \|x - \tilde{x}^*\|_2 d\varrho(x). \end{aligned}$$

Thus, for any $\tilde{r} \geq r > 0$ we obtain

$$\|x_{\alpha}^{\tilde{\mathcal{E}}}(\varrho) - \tilde{x}^*\|_2 \leq \tilde{r} + \frac{\exp\left(-\alpha\left(\inf_{x \in (B_{\tilde{r}}(\tilde{x}^*))^c} \tilde{\mathcal{E}}(x) - \tilde{\mathcal{E}}_r\right)\right)}{\varrho(B_r(\tilde{x}^*))} \int \|x - \tilde{x}^*\|_2 d\varrho(x). \quad (66)$$

We now choose $\tilde{r} = (q + \tilde{\mathcal{E}}_r)^\nu / \eta$, which satisfies $\tilde{r} \geq r$, since (64) with $\tilde{\mathcal{E}} = 0$ implies

$$\tilde{r} = \frac{(q + \tilde{\mathcal{E}}_r)^\nu}{\eta} \geq \frac{\tilde{\mathcal{E}}_r^\nu}{\eta} = \frac{\left(\sup_{x \in B_r(\tilde{x}^*)} \tilde{\mathcal{E}}(x)\right)^\nu}{\eta} \geq \sup_{x \in B_r(\tilde{x}^*)} \|x - \tilde{x}^*\|_2 = r.$$

Using again (64) with $\tilde{\mathcal{E}} = 0$ we thus have

$$\inf_{x \in (B_{\tilde{r}}(\tilde{x}^*))^c} \tilde{\mathcal{E}}(x) - \tilde{\mathcal{E}}_r \geq (\eta\tilde{r})^{1/\nu} - \tilde{\mathcal{E}}_r = q + \tilde{\mathcal{E}}_r - \tilde{\mathcal{E}}_r = q.$$

Inserting this and the definition of \tilde{r} into (66) gives the statement. \square

E.2 The auxiliary function $\tilde{\mathcal{E}}_k$

Let us now show that the function $\tilde{\mathcal{E}}_k(x) := \frac{1}{2\tau} \|x_{k-1}^{\text{CH}} - x\|_2^2 + \mathcal{E}(x)$, which appears later in the proofs of Proposition 7 and Theorem 8, satisfies the ℓ^2 -inverse continuity property in form of Definition E.1 if \mathcal{E} is Λ -convex and the parameter τ sufficiently small. As we discuss in Remark E.4 below, the condition on the parameter τ vanishes if \mathcal{E} is convex, i.e., $\Lambda \geq 0$.

Lemma E.3 ($\tilde{\mathcal{E}}_k$ satisfies the ℓ^2 -inverse continuity property). *Let $\tilde{\mathcal{E}}_k$ be defined as above with $\tau > 0$ and with $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfying A4. Moreover, if $\Lambda < 0$, assume further that $\tau < 1/(-\Lambda)$. Then, $\tilde{\mathcal{E}}_k$ satisfies the ℓ^2 -inverse continuity property (64) with parameters*

$$\nu = \frac{1}{2} \quad \text{and} \quad \eta = \sqrt{\frac{1}{2\tau} + \frac{\Lambda}{2}}.$$

I.e., denoting the unique global minimizer of $\tilde{\mathcal{E}}_k$ by \tilde{x}_k^{CH} , it holds

$$\|x - \tilde{x}_k^{\text{CH}}\|_2 \leq \frac{1}{\eta} \left(\tilde{\mathcal{E}}_k(x) - \tilde{\mathcal{E}}_k(\tilde{x}_k^{\text{CH}}) \right)^\nu \quad \text{for all } x \in \mathbb{R}^d. \quad (67)$$

Proof. We first notice that $\tilde{\mathcal{E}}_k$ is $2\eta^2 = \left(\frac{1+\Lambda\tau}{\tau}\right)$ -strongly convex ($2\eta^2 > 0$ by assumption), since

$$\begin{aligned} \tilde{\mathcal{E}}_k(x) - \frac{1}{2} \left(\frac{1+\Lambda\tau}{\tau} \right) \|x\|_2^2 &= \frac{1}{2\tau} \left(\|x_{k-1}^{\text{CH}} - x\|_2^2 - \|x\|_2^2 \right) + \mathcal{E}(x) - \frac{\Lambda}{2} \|x\|_2^2 \\ &= \underbrace{\frac{1}{2\tau} \left(\|x_{k-1}^{\text{CH}}\|_2^2 - 2\langle x_{k-1}^{\text{CH}}, x \rangle \right)}_{\text{convex since linear}} + \underbrace{\mathcal{E}(x) - \frac{\Lambda}{2} \|x\|_2^2}_{\text{convex by A4}} \end{aligned}$$

is convex by being the sum of two convex functions. By strong convexity of $\tilde{\mathcal{E}}_k$, \tilde{x}_k^{CH} exists, is unique and for all $\xi \in [0, 1]$ it holds

$$\begin{aligned} \frac{1}{2} \left(\frac{1+\Lambda\tau}{\tau} \right) \xi(1-\xi) \|x - \tilde{x}_k^{\text{CH}}\|_2^2 &\leq \xi\tilde{\mathcal{E}}_k(x) + (1-\xi)\tilde{\mathcal{E}}_k(\tilde{x}_k^{\text{CH}}) - \tilde{\mathcal{E}}_k(\xi x + (1-\xi)\tilde{x}_k^{\text{CH}}) \\ &\leq \xi \left(\tilde{\mathcal{E}}_k(x) - \tilde{\mathcal{E}}_k(\tilde{x}_k^{\text{CH}}) \right), \end{aligned}$$

where we used in the last inequality that \tilde{x}_k^{CH} minimizes $\tilde{\mathcal{E}}_k$. Dividing both sides by ξ , letting $\xi \rightarrow 0$ and reordering the inequality gives the result. \square

Remark E.4. In the case that \mathcal{E} is Λ -convex with $\Lambda < 0$ (i.e., potentially nonconvex), Lemma E.3 requires that the parameter τ is sufficiently small, in order to ensure that $\tilde{\mathcal{E}}_k$ is strongly convex and therefore has a unique global minimizer \tilde{x}_k^{CH} . On the other hand, if \mathcal{E} is convex, i.e., $\Lambda \geq 0$, $\tilde{\mathcal{E}}_k$ is strongly convex and therefore such constraint is not necessary, i.e., τ can be chosen arbitrarily.

Next, we give technical estimates on the quantities $(\tilde{\mathcal{E}}_k)_r$, $\nu_k(B_r(\tilde{x}_k^{\text{CH}}))$ and $\int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x)$, which appear when applying Proposition E.2 in the setting of the function $\tilde{\mathcal{E}}_k$ and the probability measure $\nu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, 2\tilde{\sigma}^2 \text{Id})$. This allows to keep the proof of Proposition 7 more concise.

Lemma E.5. Let $\tilde{\mathcal{E}}_k \in \mathcal{C}(\mathbb{R}^d)$ be as defined above with $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfying A2. Then for the expressions $(\tilde{\mathcal{E}}_k)_r$, $\nu_k(B_r(\tilde{x}_k^{\text{CH}}))$ and $\int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x)$ appearing in Equation (65) the following estimates hold. Namely,

$$\begin{aligned} (\tilde{\mathcal{E}}_k)_r &\leq \left(\frac{1}{2\tau} \left(r + 4\tau C_1 (\|x_{k-1}^{\text{CH}}\|_2 + \|\tilde{x}_k^{\text{CH}}\|_2) \right) + C_1 (r + 2\|\tilde{x}_k^{\text{CH}}\|_2) \right) r, \\ \nu_k(B_r(\tilde{x}_k^{\text{CH}})) &\geq \frac{1}{(2\tilde{\sigma})^d} \exp \left(-\frac{1}{2\tilde{\sigma}^2} \left(r^2 + 8\tau^2 C_1^2 (\|x_{k-1}^{\text{CH}}\|_2^2 + \|\tilde{x}_k^{\text{CH}}\|_2^2) \right) \right) \frac{1}{\Gamma(\frac{d}{2} + 1)} r^d, \\ \int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x) &\leq 2\tau C_1 (\|x_{k-1}^{\text{CH}}\|_2 + \|\tilde{x}_k^{\text{CH}}\|_2) + \sqrt{2d}\tilde{\sigma}. \end{aligned}$$

Proof. Let us start by investigating the expressions $(\tilde{\mathcal{E}}_k)_r$, $\nu_k(B_r(\tilde{x}_k^{\text{CH}}))$ and $\int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x)$ individually.

Term $(\tilde{\mathcal{E}}_k)_r$: By definition (see Proposition E.2) and under A2 it holds

$$\begin{aligned} (\tilde{\mathcal{E}}_k)_r &= \sup_{x \in B_r(\tilde{x}_k^{\text{CH}})} \tilde{\mathcal{E}}_k(x) - \tilde{\mathcal{E}}_k(\tilde{x}_k^{\text{CH}}) \\ &\leq \frac{1}{2\tau} \sup_{x \in B_r(\tilde{x}_k^{\text{CH}})} \left(\|x_{k-1}^{\text{CH}} - x\|_2^2 - \|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \right) + \sup_{x \in B_r(\tilde{x}_k^{\text{CH}})} \mathcal{E}(x) - \mathcal{E}(\tilde{x}_k^{\text{CH}}) \\ &\leq \frac{1}{2\tau} (r + 2\|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2) r + C_1 (r + 2\|\tilde{x}_k^{\text{CH}}\|_2) r \\ &\leq \left(\frac{1}{2\tau} (r + 2\|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2) + C_1 (r + 2\|\tilde{x}_k^{\text{CH}}\|_2) \right) r. \end{aligned}$$

Term $\nu_k(B_r(\tilde{x}_k^{\text{CH}}))$: Using the density of the multivariate normal distribution $\nu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, 2\tilde{\sigma}^2 \text{Id})$ we can directly compute

$$\begin{aligned} \nu_k(B_r(\tilde{x}_k^{\text{CH}})) &= \frac{1}{(4\pi\tilde{\sigma}^2)^{d/2}} \int_{B_r(\tilde{x}_k^{\text{CH}})} \exp \left(-\frac{1}{4\tilde{\sigma}^2} \|x - x_{k-1}^{\text{CH}}\|_2^2 \right) d\lambda(x) \\ &\geq \frac{1}{(4\pi\tilde{\sigma}^2)^{d/2}} \int_{B_r(\tilde{x}_k^{\text{CH}})} \exp \left(-\frac{1}{2\tilde{\sigma}^2} \left(\|x - \tilde{x}_k^{\text{CH}}\|_2^2 + \|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2^2 \right) \right) d\lambda(x) \\ &\geq \frac{1}{(4\pi\tilde{\sigma}^2)^{d/2}} \exp \left(-\frac{1}{2\tilde{\sigma}^2} \left(r^2 + \|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2^2 \right) \right) \int_{B_r(\tilde{x}_k^{\text{CH}})} d\lambda(x) \\ &= \frac{1}{(2\tilde{\sigma})^d} \exp \left(-\frac{1}{2\tilde{\sigma}^2} \left(r^2 + \|\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}\|_2^2 \right) \right) \frac{1}{\Gamma(\frac{d}{2} + 1)} r^d, \end{aligned}$$

where we used in the last step that the volume of a d -dimensional unit ball is $\pi^{d/2}/\Gamma(\frac{d}{2} + 1)$. Here, Γ denotes Euler's gamma function. We recall for the readers' convenience that by Stirling's approximation $\Gamma(x+1) \sim \sqrt{2\pi x} (x/e)^x$ as $x \rightarrow \infty$.

Term $\int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x)$: A straightforward computation gives

$$\begin{aligned} \int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x) &= \int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\mathcal{N}(x_{k-1}^{\text{CH}}, 2\tilde{\sigma}^2 \text{Id})(x) \\ &= \int \|x + x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2 d\mathcal{N}(0, 2\tilde{\sigma}^2 \text{Id})(x) \\ &\leq \|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2 + \int \|x\|_2 d\mathcal{N}(0, 2\tilde{\sigma}^2 \text{Id})(x) \\ &\leq \|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2 + \sqrt{2d}\tilde{\sigma}. \end{aligned}$$

Concluding step: To conclude the proof, we further observe that since \tilde{x}_k^{CH} is the minimizer of $\tilde{\mathcal{E}}_k$, see (14), a comparison with x_{k-1}^{CH} yields

$$\frac{1}{2\tau} \|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 + \mathcal{E}(\tilde{x}_k^{\text{CH}}) \leq \mathcal{E}(x_{k-1}^{\text{CH}}).$$

With A2 it therefore holds

$$\|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \leq 2\tau (\mathcal{E}(x_{k-1}^{\text{CH}}) - \mathcal{E}(\tilde{x}_k^{\text{CH}})) \leq 2\tau C_1 (\|x_{k-1}^{\text{CH}}\|_2 + \|\tilde{x}_k^{\text{CH}}\|_2) \|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2,$$

or rephrased

$$\|x_{k-1}^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2 \leq 2\tau C_1 (\|x_{k-1}^{\text{CH}}\|_2 + \|\tilde{x}_k^{\text{CH}}\|_2).$$

Exploiting this estimate in the former bounds, gives the statements. \square

E.3 Proof of Proposition 7

We now have all necessary tools at hand to present the detailed proof of Proposition 7.

Proof of Proposition 7. By using the quantitative Laplace principle E.2, we make rigorous and quantify the fact that x_k^{CH} approximates the minimizer of $\tilde{\mathcal{E}}_k$, denoted by \tilde{x}_k , for sufficiently large α .

To obtain the probabilistic formulation of the statement, let us again denote the underlying probability space by $(\Omega, \mathcal{F}, \mathbb{P})$ (note that we can use the same probability space as in Section D since the stochasticity of both schemes (13) and (14) is solely coming from the initialization) and introduce the subset $\tilde{\Omega}_M$ of Ω of suitably bounded random variables according to

$$\tilde{\Omega}_M := \left\{ \omega \in \Omega : \max_{k=0, \dots, K} \max \{ \|x_k^{\text{CH}}\|_2, \|\tilde{x}_k^{\text{CH}}\|_2 \} \leq M \right\}.$$

For the associated cutoff function (random variable) we write $\mathbb{1}_{\tilde{\Omega}_M}$.

We first notice that by definition of the consensus point $x_\alpha^\mathcal{E}$ in (3) it holds

$$\begin{aligned} x_k^{\text{CH}} = x_\alpha^\mathcal{E}(\mu_k) &= \int x \frac{\exp(-\alpha\mathcal{E}(x))}{\|\exp(-\alpha\mathcal{E})\|_{L^1(\mu_k)}} d\mu_k(x) \\ &= \int x \frac{\exp(-\alpha\mathcal{E}(x)) \exp\left(-\frac{1}{4\tilde{\sigma}^2} \|x - x_{k-1}^{\text{CH}}\|_2^2\right)}{\int \exp(-\alpha\mathcal{E}(x')) \exp\left(-\frac{1}{4\tilde{\sigma}^2} \|x' - x_{k-1}^{\text{CH}}\|_2^2\right) d\nu_k(x')} d\nu_k(x) \\ &= \int x \frac{\exp(-\alpha\tilde{\mathcal{E}}_k(x))}{\|\exp(-\alpha\tilde{\mathcal{E}}_k)\|_{L^1(\nu_k)}} d\nu_k(x) \\ &= x_\alpha^{\tilde{\mathcal{E}}_k}(\nu_k), \end{aligned} \tag{68}$$

which introduces the relation $\tau = 2\alpha\tilde{\sigma}^2$ and where we chose $\nu_k = \mathcal{N}(x_{k-1}^{\text{CH}}, 2\tilde{\sigma}^2 \text{Id})$, which is globally supported, i.e., $\text{supp}(\nu_k) = \mathbb{R}^d$. Since, according to Lemma E.3, $\tilde{\mathcal{E}}_k$ satisfies the inverse

continuity property (67) with $\nu = 1/2$ and $\eta = \sqrt{\frac{1}{2\tau} + \frac{\Lambda}{2}} > 0$, the quantitative Laplace principle, Proposition E.2, gives for any $r, q > 0$ the bound

$$\|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2 = \|x_{\alpha}^{\tilde{\mathcal{E}}_k}(\nu_k) - \tilde{x}_k^{\text{CH}}\|_2 \leq \frac{(q + (\tilde{\mathcal{E}}_k)_r)^\nu}{\eta} + \frac{\exp(-\alpha q)}{\nu_k(B_r(\tilde{x}_k^{\text{CH}}))} \int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x), \quad (69)$$

where $(\tilde{\mathcal{E}}_k)_r := \sup_{x \in B_r(\tilde{x}_k^{\text{CH}})} \tilde{\mathcal{E}}_k(x) - \tilde{\mathcal{E}}_k(\tilde{x}_k^{\text{CH}})$. We further notice that by the assumption $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$ it holds $\eta \geq 1/(2\sqrt{\tau})$ (in the case $\Lambda \geq 0$ the same bound holds trivially). Combining (69) with the technical estimates of Lemma E.5 and the definition of the cutoff function $\mathbb{1}_{\tilde{\Omega}_M}$ allows to obtain

$$\begin{aligned} & \mathbb{E} \|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \\ & \leq 2\mathbb{E} \left[\frac{(q + (\tilde{\mathcal{E}}_k)_r)}{\eta^2} \mathbb{1}_{\tilde{\Omega}_M} \right] + 2\mathbb{E} \left[\frac{\exp(-2\alpha q)}{\nu_k(B_r(\tilde{x}_k^{\text{CH}}))^2} \left(\int \|x - \tilde{x}_k^{\text{CH}}\|_2 d\nu_k(x) \right)^2 \mathbb{1}_{\tilde{\Omega}_M} \right] \\ & \leq 8\tau \left(q + \left(\frac{r}{2\tau} + C_1 r + 6C_1 M \right) r \right) \\ & \quad + 4 \exp \left(-2\alpha q + \frac{1}{\tilde{\sigma}^2} (r^2 + 16\tau^2 C_1^2 M^2) \right) \frac{2^d (2\tilde{\sigma}^2)^d}{r^{2d}} \Gamma \left(\frac{d}{2} + 1 \right)^2 (16\tau^2 C_1^2 M^2 + 2d\tilde{\sigma}^2) \\ & = 8\tau \left(q + \left(\frac{r}{2\tau} + C_1 r + 6C_1 M \right) r \right) \\ & \quad + 4 \exp \left(-2\alpha \left(q - \left(\frac{r^2}{\tau} + 16\tau C_1^2 M^2 \right) \right) \right) \frac{2^d \tau^d}{\alpha^d r^{2d}} \Gamma \left(\frac{d}{2} + 1 \right)^2 \left(16\tau^2 C_1^2 M^2 + d \frac{\tau}{\alpha} \right), \end{aligned} \quad (70)$$

where in the last step we just replaced $2\tilde{\sigma}^2$ by τ/α according to the relation. We now choose

$$r = \tau, \quad q = \frac{3}{2}\tau + 16\tau C_1^2 M^2 \quad \text{and} \quad \alpha \geq \alpha_0 := \frac{1}{\tau} \left(d \log 2 + \log(1+d) + 2 \log \Gamma \left(\frac{d}{2} + 1 \right) \right),$$

where Γ denotes Euler's gamma function, for which, by Stirling's approximation, it holds $\Gamma(x+1) \sim \sqrt{2\pi x} (x/e)^x$ as $x \rightarrow \infty$. With this we can continue the computations of (70) with

$$\begin{aligned} \mathbb{E} \|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} & \leq 8 \left(2 + C_1 \tau + 6C_1 M + 16C_1^2 M^2 \right) \tau^2 \\ & \quad + 4 \exp(-\alpha\tau) \frac{2^d}{\alpha^d \tau^d} \Gamma \left(\frac{d}{2} + 1 \right)^2 \left(16\tau^2 C_1^2 M^2 + d \frac{\tau}{\alpha} \right) \\ & \leq 8 \left(3 + C_1 \tau + 6C_1 M + 24C_1^2 M^2 \right) \tau^2 \\ & \leq c\tau^2 \end{aligned} \quad (71)$$

with a constant $c = c(C_1, M)$. Notice that to obtain the last inequality one may first note and exploit that one has $\alpha\tau \geq 1$ as well as $1/\alpha \leq \tau$ as a consequence of $\alpha \geq 1/\tau$.

Probabilistic formulation: We first note that with Markov's inequality we have the estimate

$$\begin{aligned} \mathbb{P}(\tilde{\Omega}_M^c) & = \mathbb{P} \left(\max_{k=0, \dots, K} \max \{ \|x_k^{\text{CH}}\|_2, \|\tilde{x}_k^{\text{CH}}\|_2 \} > M \right) \\ & \leq \frac{1}{M^4} \left(\mathbb{E} \max_{k=0, \dots, K} \|x_k^{\text{CH}}\|_2^4 + \mathbb{E} \max_{k=0, \dots, K} \|\tilde{x}_k^{\text{CH}}\|_2^4 \right) \\ & \leq \frac{1}{M^4} (\mathcal{M}^{\text{CH}} + \tilde{\mathcal{M}}^{\text{CH}}), \end{aligned}$$

where the last inequality is due to Lemmas C.3 and C.6. Thus, for any $\delta \in (0, 1/2)$, a sufficiently large choice $M = M(\delta^{-1}, \mathcal{M}^{\text{CH}}, \tilde{\mathcal{M}}^{\text{CH}})$ allows to ensure $\mathbb{P}(\tilde{\Omega}_M^c) \leq \delta$. To conclude the proof, let us denote by $\tilde{K}_\varepsilon \subset \Omega$ the set, where (17) does not hold and abbreviate

$$\epsilon = \varepsilon^{-1} c\tau^2.$$

For the probability of this set we can estimate

$$\begin{aligned}\mathbb{P}(\tilde{K}_\varepsilon) &= \mathbb{P}(\tilde{K}_\varepsilon \cap \tilde{\Omega}_M) + \mathbb{P}(\tilde{K}_\varepsilon \cap \tilde{\Omega}_M^c) \leq \mathbb{P}(\tilde{K}_\varepsilon \mid \tilde{\Omega}_M) \mathbb{P}(\tilde{\Omega}_M) + \mathbb{P}(\tilde{\Omega}_M^c) \\ &\leq \mathbb{P}(\tilde{K}_\varepsilon \mid \tilde{\Omega}_M) + \delta \leq \epsilon^{-1} \mathbb{E} \left[\|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mid \tilde{\Omega}_M \right] + \delta,\end{aligned}\quad (72)$$

where the last step is due to Markov's inequality. By definition of the conditional expectation we further have

$$\mathbb{E} \left[\|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mid \tilde{\Omega}_M \right] \leq \frac{1}{\mathbb{P}(\tilde{\Omega}_M)} \mathbb{E} \|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \leq 2 \mathbb{E} \|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M}.$$

Inserting now the expression from (71) concludes the proof. \square

E.4 Proof of Theorem 8

We now have all necessary tools at hand to present the detailed proof of Theorem 8.

Proof of Theorem 8. We combine in what follows Proposition 7 with a stability argument for the MMS (15).

To obtain the probabilistic formulation of the statement, let us denote, as in the proof of Proposition 7, the underlying probability space by $(\Omega, \mathcal{F}, \mathbb{P})$ (note that we can use the same probability space as in Section D since the stochasticity of the three schemes (13), (14) and (15) is solely coming from the initialization) and introduce the subset $\tilde{\Omega}_M$ of Ω of suitably bounded random variables according to

$$\tilde{\Omega}_M := \left\{ \omega \in \Omega : \max_{k=0, \dots, K} \max \{ \|x_k^{\text{CH}}\|_2, \|\tilde{x}_k^{\text{CH}}\|_2 \} \leq M \right\}.$$

For the associated cutoff function (random variable) we write $\mathbb{1}_{\tilde{\Omega}_M}$.

We can decompose the expected squared discrepancy $\mathbb{E} \|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M}$ between the MMS (15) and the CH scheme (13) for any $\vartheta \in (0, 1)$ as

$$\mathbb{E} \|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \leq (1 + \vartheta) \mathbb{E} \|x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} + (1 + \vartheta^{-1}) \mathbb{E} \|\tilde{x}_k^{\text{CH}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M}.\quad (73)$$

In what follows we individually estimate the two terms on the right-hand side of (73).

First term: Let us first bound the term $\mathbb{E} \|x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M}$. By definition of x_k^{MMS} and \tilde{x}_k^{CH} as minimizers of (15) and (14), respectively, and with the definition $\mathcal{E}_\Lambda(x) := \mathcal{E}(x) - \frac{\Lambda}{2} \|x\|_2^2$ it holds

$$\frac{(1 + \tau\Lambda)x_k^{\text{MMS}} - x_{k-1}^{\text{MMS}}}{\tau} \in -\partial\mathcal{E}_\Lambda(x_k^{\text{MMS}}) \quad \text{and} \quad \frac{(1 + \tau\Lambda)\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}}{\tau} \in -\partial\mathcal{E}_\Lambda(\tilde{x}_k^{\text{CH}}).$$

Since \mathcal{E}_Λ is convex due to A4 and as consequence of the properties of the subdifferential we have

$$\left\langle -\frac{(1 + \tau\Lambda)x_k^{\text{MMS}} - x_{k-1}^{\text{MMS}}}{\tau} + \frac{(1 + \tau\Lambda)\tilde{x}_k^{\text{CH}} - x_{k-1}^{\text{CH}}}{\tau}, x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}} \right\rangle \geq 0,$$

which allows to obtain by means of Cauchy-Schwarz inequality

$$(1 + \tau\Lambda) \|x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}}\|_2^2 \leq \langle x_{k-1}^{\text{MMS}} - x_{k-1}^{\text{CH}}, x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}} \rangle \leq \|x_{k-1}^{\text{MMS}} - x_{k-1}^{\text{CH}}\|_2 \|x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}}\|_2$$

or, equivalently,

$$\|x_k^{\text{MMS}} - \tilde{x}_k^{\text{CH}}\|_2 \leq \frac{1}{1 + \tau\Lambda} \|x_{k-1}^{\text{MMS}} - x_{k-1}^{\text{CH}}\|_2.\quad (74)$$

Second term: For the term $\mathbb{E} \|\tilde{x}_k^{\text{CH}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M}$ we obtained in (71) in the proof of Proposition 7, for suitable choices of $\tilde{\sigma}$ and α , the bound

$$\mathbb{E} \|x_k^{\text{CH}} - \tilde{x}_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \leq c\tau^2\quad (75)$$

with a constant $c = c(C_1, M)$.

Concluding step: Combining this with the estimate (74) yields for (73) the bound

$$\mathbb{E} \|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \leq \frac{1 + \vartheta}{(1 + \tau\Lambda)^2} \mathbb{E} \|x_{k-1}^{\text{MMS}} - x_{k-1}^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} + c(1 + \vartheta^{-1}) \tau^2. \quad (76)$$

An application of the discrete variant of Grönwall's inequality (22) shows that

$$\mathbb{E} \|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \leq c(1 + \vartheta^{-1}) \tau^2 \sum_{\ell=0}^{k-1} \left(\frac{1 + \vartheta}{(1 + \tau\Lambda)^2} \right)^\ell \quad (77)$$

for all $k = 1, \dots, K$, where we used that both schemes are initialized by the same x_0 .

Probabilistic formulation: We first note that with Markov's inequality we have the estimate

$$\begin{aligned} \mathbb{P}(\tilde{\Omega}_M^c) &= \mathbb{P} \left(\max_{k=0, \dots, K} \max \{ \|x_k^{\text{CH}}\|_2, \|\tilde{x}_k^{\text{CH}}\|_2 \} > M \right) \\ &\leq \frac{1}{M^4} \left(\mathbb{E} \max_{k=0, \dots, K} \|x_k^{\text{CH}}\|_2^4 + \mathbb{E} \max_{k=0, \dots, K} \|\tilde{x}_k^{\text{CH}}\|_2^4 \right) \\ &\leq \frac{1}{M^4} (\mathcal{M}^{\text{CH}} + \tilde{\mathcal{M}}^{\text{CH}}), \end{aligned}$$

where the last inequality is due to Lemmas C.3 and C.6. Thus, for any $\delta \in (0, 1/2)$, a sufficiently large choice $M = M(\delta^{-1}, \mathcal{M}^{\text{CH}}, \tilde{\mathcal{M}}^{\text{CH}})$ allows to ensure $\mathbb{P}(\tilde{\Omega}_M^c) \leq \delta$. To conclude the proof, let us denote by $\tilde{K}_\varepsilon \subset \Omega$ the set, where (18) does not hold and abbreviate

$$\varepsilon = \varepsilon^{-1} c(1 + \vartheta^{-1}) \tau^2 \sum_{\ell=0}^{k-1} \left(\frac{1 + \vartheta}{(1 + \tau\Lambda)^2} \right)^\ell.$$

For the probability of this set we can estimate

$$\begin{aligned} \mathbb{P}(\tilde{K}_\varepsilon) &= \mathbb{P}(\tilde{K}_\varepsilon \cap \tilde{\Omega}_M) + \mathbb{P}(\tilde{K}_\varepsilon \cap \tilde{\Omega}_M^c) \leq \mathbb{P}(\tilde{K}_\varepsilon \mid \tilde{\Omega}_M) \mathbb{P}(\tilde{\Omega}_M) + \mathbb{P}(\tilde{\Omega}_M^c) \\ &\leq \mathbb{P}(\tilde{K}_\varepsilon \mid \tilde{\Omega}_M) + \delta \leq \varepsilon^{-1} \mathbb{E} \left[\|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mid \tilde{\Omega}_M \right] + \delta, \end{aligned} \quad (78)$$

where the last step is due to Markov's inequality. By definition of the conditional expectation we further have

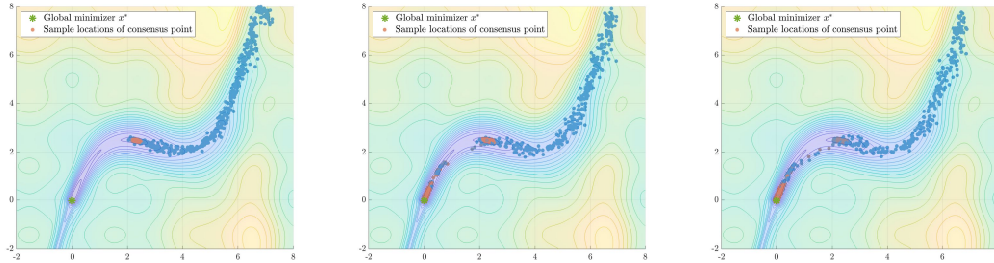
$$\mathbb{E} \left[\|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mid \tilde{\Omega}_M \right] \leq \frac{1}{\mathbb{P}(\tilde{\Omega}_M)} \mathbb{E} \|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M} \leq 2 \mathbb{E} \|x_k^{\text{MMS}} - x_k^{\text{CH}}\|_2^2 \mathbb{1}_{\tilde{\Omega}_M}.$$

Inserting now the expression from (77) concludes the proof. \square

F Additional numerical experiments

F.1 Comparison of the CH scheme (13) for different sampling widths $\tilde{\sigma}$

To complement Figure 2a, we visualize in Figure F.1 the influence of the sampling width $\tilde{\sigma}$ on the behavior of the CH scheme (13).



(a) The CH scheme (13) with sampling width $\tilde{\sigma} = 0.4$ gets stuck in a local minimum of \mathcal{E} .

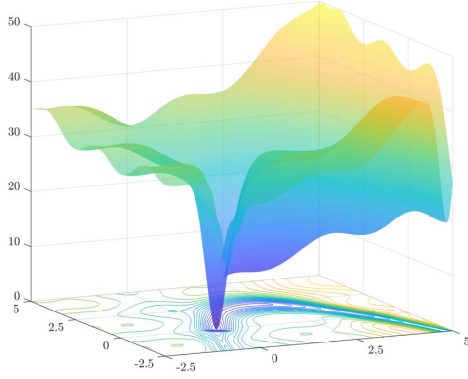
(b) The CH scheme (13) with sampling width $\tilde{\sigma} = 0.6$ can occasionally escape local minima of \mathcal{E} .

(c) The CH scheme (13) with sampling width $\tilde{\sigma} = 0.7$ can escape local minima of \mathcal{E} .

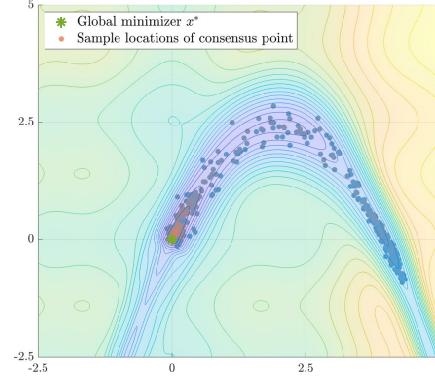
Figure F.1: A visual comparison of the CH scheme (13) for different sampling widths $\tilde{\sigma}$. We depict the positions of the consensus hopping scheme (13) for different values of $\tilde{\sigma}$ (0.4 in (a), 0.6 in (b) and 0.7 in (c)) in the setting of Figure 2a. While for small $\tilde{\sigma}$ the numerical scheme gets stuck in a local minimum of the objective, the ability to escape such critical points improves with larger $\tilde{\sigma}$. Notice that (b) coincides with Figure 2a.

F.2 The numerical experiments of Figures 1 and 2 for a different objective

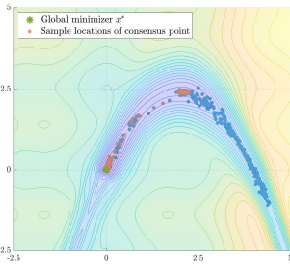
In the style of Figures 1 and 2 we provide in Figure F.2 an additional set of illustrations of the behavior of the different algorithms analyzed in this work for a noisy Canyon function with a valley shaped as a second degree polynomial.



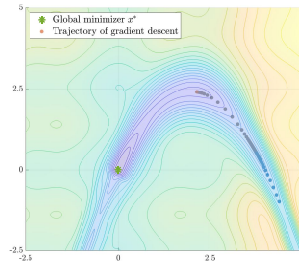
(a) A noisy Canyon function \mathcal{E} with a valley shaped as a second degree polynomial



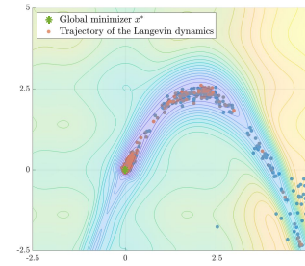
(b) The CBO scheme (4) (sampled over several runs) follows on average the valley while passing over local minima.



(c) The CH scheme (13) (sampled over several runs) follows on average the valley of \mathcal{E} and can occasionally escape local minima.



(d) Gradient descent gets stuck in a local minimum of \mathcal{E} .



(e) The Langevin dynamics (6) (sampled over several runs) follows on average the valley of \mathcal{E} and escapes local minima.

Figure F.2: An additional numerical experiment illustrating the behavior of the CBO scheme (4) (see (b)), the consensus hopping scheme (13) (see (c)), gradient descent (see (d)) and the overdamped Langevin dynamics (6) (see (e)) in search of the global minimizer x^* of the nonconvex objective function \mathcal{E} depicted in (a). The experimental setting is the one of Figures 1 and 2 with the only difference of the particles being initialized around $(5, -1)$.

License for [CBO&GD].

The permission to reprint and include the material is printed on the next page(s).

PMLR Proceedings of Machine Learning Research

[\[edit\]](#)

Proceedings of Machine Learning Research

ISSN: 2640-3498

The Proceedings of Machine Learning Research is a series that publishes machine learning research papers presented at workshops and conferences. Each volume is separately titled and associated with a particular workshop or conference and will be published online on the PMLR web site. Authors retain copyright.

Editors

The Series Editor is [Neil Lawrence](#). Please send proposals for new volumes under this series to us via e-mail: proceedings@mlr.press. Each proposal should include:

- A brief description of the event's scope and topics to be covered.
- A description of the review process for the proceedings.
- The names and short CVs (a few lines) of the proposed Proceedings Editors.

For frequently asked questions on publishing proceedings please see the [FAQ](#).

For details on how to prepare a proceedings for publication please see the [Specification](#).

arXiv.org - Non-exclusive license to distribute

The URI <http://arxiv.org/licenses/nonexclusive-distrib/1.0/> is used to record the fact that the submitter granted the following license to arXiv.org on submission of an article:

- I grant arXiv.org a perpetual, non-exclusive license to distribute this article.
- I certify that I have the right to grant this license.
- I understand that submissions cannot be completely removed once accepted.
- I understand that arXiv.org reserves the right to reclassify or reject any submission.

Revision history

2004-01-16 - License above introduced as part of arXiv submission process

2007-06-21 - This HTML page created

[Contact](#)

Paper P7

On the Global Convergence of Particle Swarm Optimization Methods

H. Huang, J. Qiu, and K. Riedl
Appl. Math. Optim. (2023)

Paper Summary of [PSO]⁴⁰

In the paper “On the Global Convergence of Particle Swarm Optimization Methods,” published in the *Applied Mathematics and Optimization Journal*, we prove the convergence of PSO methods (5.1) to global minimizers.

PSO, originally proposed by the authors of [KE95; Ken97], is a renowned, well-known, and widely employed metaheuristic optimization algorithm suitable for solving nonconvex nonsmooth problems of the form (2.1). Its design is inspired by the fascinating capabilities of swarm intelligence observed in nature.

Employing tools from stochastic calculus and the analysis of partial differential equations, we provide in [PSO] a rigorous convergence analysis to global minimizers for PSO. We model the PSO dynamics as suggested in [GP21], where a continuous description based on a system of SDEs is provided. Our analysis follows the framework put forward in the line of works [Car+18; Car+21], i.e, we establish convergence to a global minimizer of a possibly nonconvex and nonsmooth objective function in two steps. First, we prove consensus formation of an associated mean-field dynamics by analyzing the time-evolution of the variance of the particle distribution, which acts as Lyapunov function of the dynamics. We then show that this consensus is close to a global minimizer by employing the asymptotic Laplace principle and a tractability condition on the energy landscape of the objective function. These results allow for the usage of memory mechanisms, and hold for a rich class of objectives provided certain conditions of well-preparation of the hyperparameters and the initial datum [PSO, Sections 2 and 3]. In a second step, at least for the case without memory effects, we provide a quantitative result about the mean-field approximation of particle swarm optimization, which specifies the convergence of the interacting particle system to the associated mean-field limit. Combining these two results allows for global convergence guarantees of the numerical particle swarm optimization method with provable polynomial complexity [PSO, Section 4]. To demonstrate the applicability of the method we propose an efficient and parallelizable implementation, which is tested in particular on a competitive and well-understood high-dimensional benchmark problem in machine learning [PSO, Section 5].

KR’s Contributions. HH and JQ suggested to extend the convergence analysis framework developed originally for CBO to the well-known PSO method. Together with HH, KR worked out the convergence proof for the mean-field limit of PSO, both in the setting with and without memory effects, and eventually proving convergence to global minimizers under suitable well-preparedness assumptions. For PSO without memory effects, KR further devised a mean-field approximation, allowing for a holistic convergence statement. KR conducted the numerical experiments, coded an efficient implementation of PSO using random mini-batch ideas as well as traditional metaheuristic-inspired techniques from genetic programming and simulated annealing, and wrote large parts of the paper, which was proofread and refined by JQ and HH.

⁴⁰In this section, we follow [PSO, Abstract].


The following document is a reprint of

[PSO] H. Huang, J. Qiu, and K. Riedl. “On the Global Convergence of Particle Swarm Optimization Methods.” In: *Appl. Math. Optim.* 88.2 (2023), Paper No. 30, 44.

The permission to reprint and include the material is provided after the reprint.



On the Global Convergence of Particle Swarm Optimization Methods

Hui Huang¹ · Jinniao Qiu² · Konstantin Riedl^{3,4} 

Accepted: 26 February 2023 / Published online: 31 May 2023
© The Author(s) 2023

Abstract

In this paper we provide a rigorous convergence analysis for the renowned particle swarm optimization method by using tools from stochastic calculus and the analysis of partial differential equations. Based on a continuous-time formulation of the particle dynamics as a system of stochastic differential equations, we establish convergence to a global minimizer of a possibly nonconvex and nonsmooth objective function in two steps. First, we prove consensus formation of an associated mean-field dynamics by analyzing the time-evolution of the variance of the particle distribution, which acts as Lyapunov function of the dynamics. We then show that this consensus is close to a global minimizer by employing the asymptotic Laplace principle and a tractability condition on the energy landscape of the objective function. These results allow for the usage of memory mechanisms, and hold for a rich class of objectives provided certain conditions of well-preparation of the hyperparameters and the initial datum. In a second step, at least for the case without memory effects, we provide a quantitative result about the mean-field approximation of particle swarm optimization, which specifies the convergence of the interacting particle system to the associated mean-field limit. Combining these two results allows for global convergence guarantees of the numerical particle swarm optimization method with provable polynomial complexity. To demonstrate the applicability of the method we propose an efficient and parallelizable

✉ Konstantin Riedl
konstantin.riedl@ma.tum.de

Hui Huang
hui.huang@uni-graz.at

Jinniao Qiu
jinniao.qiu@ucalgary.ca

- ¹ Institute of Mathematics and Scientific Computing, University of Graz, Graz, Austria
- ² Department of Mathematics and Statistics, University of Calgary, Calgary, Canada
- ³ Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
- ⁴ Munich Center for Machine Learning, Munich, Germany

implementation, which is tested in particular on a competitive and well-understood high-dimensional benchmark problem in machine learning.

Keywords Global derivative-free optimization · High-dimensional nonconvex optimization · Metaheuristics · Particle swarm optimization · Mean-field limit · Vlasov-Fokker-Planck equations

Mathematics Subject Classification 65C35 · 65K10 · 90C26 · 90C56 · 35Q90 · 35Q83

1 Introduction

In nature, collective behavior and self-organization allow complicated global patterns to emerge from simple interaction rules and random fluctuations. Inspired by the fascinating capabilities of swarm intelligence, large multi-agent systems are employed as a tool for solving challenging problems in applied mathematics. One classical task arising throughout science is concerned with the global optimization of a problem-dependent possibly nonconvex and nonsmooth objective function $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e., the search for a global optimizer

$$x^* \in \arg \min_{x \in \mathbb{R}^d} \mathcal{E}(x). \quad (1.1)$$

A popular class of methods with a long history of achieving state-of-the-art performance on such problems are metaheuristics [14]. They orchestrate an interplay between local and global improvement procedures, consider memory mechanisms and selection strategies, and combine random and deterministic decisions, to create a process capable of escaping local optima and performing a robust search of the solution space in order to find a global optimizer. Initiated by seminal works on stochastic approximation [49] and random search [46], a big variety of such mechanisms has been introduced, analyzed and applied to numerous real-world problems. A non-exclusive list of representatives includes evolutionary programming [15], genetic algorithms [24], simulated annealing [1], and particle swarm optimization [31]. Despite their tremendous empirical success, it is very difficult to provide a theoretical convergence analysis to global minimizers, mostly due to their stochastic nature and the appearance of memory effects. Simulated annealing, however, is theoretically actually well-studied, see, e.g., the works [3, 37] as well as the recent survey [53] and references therein.

In this paper we study particle swarm optimization (PSO), which was initially introduced by Kennedy and Eberhart in the 90s [30, 31] and is now widely recognized as an efficient method for tackling complex optimization problems [35, 45]. Originally, PSO solves (1.1) by considering a group of finitely many particles, which explore the energy landscape of \mathcal{E} . Each agent experiences a force towards its own personal (historical) best position as well as towards the global best position communicated in the swarm. We refer to the ability of each particle remembering the best position it has been positioned at in the past as memory mechanisms. Although these interaction rules

are seemingly simple, a complete numerical analysis of PSO is still lacking; see, e.g., [41, 55, 57] and references therein. Recently, however, by introducing a continuous description of PSO based on a system of stochastic differential equations (SDEs), the authors of [22] have paved the way for a rigorous mathematical analysis using tools from stochastic calculus and the analysis of partial differential equations (PDEs).

In order to explore the domain and to form a global consensus about the minimizer x^* as time passes, the formulation of PSO proposed by the authors of [22] uses N particles, described by triplets $((X_t^i, Y_t^i, V_t^i)_{t \geq 0})_{i=1, \dots, N}$, with X_t^i and V_t^i denoting the position and velocity, and Y_t^i being a regularized version of the local (historical) best position, also referred to as memory, of the i th agent at time t . The particles, formally stochastic processes, are initialized independently according to some common distribution $f_0 \in \mathcal{P}(\mathbb{R}^{3d})$. In the most general form the PSO dynamics is given by the system of SDEs, expressed in Itô's form as

$$dX_t^i = V_t^i dt, \tag{1.2a}$$

$$dY_t^i = \kappa \left(X_t^i - Y_t^i \right) S^{\beta, \theta} \left(X_t^i, Y_t^i \right) dt, \tag{1.2b}$$

$$m dV_t^i = -\gamma V_t^i dt + \lambda_1 \left(Y_t^i - X_t^i \right) dt + \lambda_2 \left(y_\alpha \left(\widehat{\rho}_{Y,t}^N \right) - X_t^i \right) dt + \sigma_1 D \left(Y_t^i - X_t^i \right) dB_t^{1,i} + \sigma_2 D \left(y_\alpha \left(\widehat{\rho}_{Y,t}^N \right) - X_t^i \right) dB_t^{2,i}, \tag{1.2c}$$

where $\alpha, \beta, \theta, \kappa, \gamma, m, \lambda_1, \lambda_2, \sigma_1, \sigma_2 \geq 0$ are user-specified parameters. The change of the velocity in (1.2c) is subject to five forces. The first term on the right-hand side models friction with a coefficient commonly chosen as $\gamma = 1 - m \geq 0$, where $m > 0$ denotes the inertia weight. The subsequent term can be regarded as the drift towards the local best position of the i th particle, which it has memorized in the state variable Y_t^i . A continuous-time approximation of its evolution is given by Y_t^i and described in Equation (1.2b). It involves the operator $S^{\beta, \theta}$, given by $S^{\beta, \theta}(x, y) = 1 + \theta + \tanh(\beta(\mathcal{E}(y) - \mathcal{E}(x)))$ for $0 \leq \theta \ll 1$ and $\beta \gg 1$, which converges to the Heaviside function as $\theta \rightarrow 0$ and $\beta \rightarrow \infty$. The concept behind Equation (1.2b) can then be seen when being discretized, see Remark 1. For an alternative implementation of the local best position we refer to [54].

Remark 1 A time-discretization of (1.2b) with $\kappa = 1/(2\Delta t)$, $\theta = 0$ and $\beta = \infty$ yields the update rule

$$Y_{(k+1)\Delta t}^i = \begin{cases} Y_{k\Delta t}^i, & \text{if } \mathcal{E}(X_{(k+1)\Delta t}^i) \geq \mathcal{E}(Y_{k\Delta t}^i), \\ X_{(k+1)\Delta t}^i, & \text{if } \mathcal{E}(X_{(k+1)\Delta t}^i) < \mathcal{E}(Y_{k\Delta t}^i), \end{cases} \tag{1.3}$$

meaning that the i th particle stores in $Y_{k\Delta t}^i$ the best position which it has seen up to the k th iteration. This explains the name local (historical) best position and restores the original definition from the work [31].

The last deterministic term imposes a drift towards the momentaneous consensus point $y_\alpha(\widehat{\rho}_{Y,t}^N)$, given by

$$y_\alpha(\widehat{\rho}_{Y,t}^N) := \int_{\mathbb{R}^d} y \frac{\omega_\alpha^\mathcal{E}(y)}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\widehat{\rho}_{Y,t}^N)}} d\widehat{\rho}_{Y,t}^N(y), \quad \text{with} \quad \omega_\alpha^\mathcal{E}(y) := \exp(-\alpha\mathcal{E}(y)), \quad (1.4)$$

where $\widehat{\rho}_{Y,t}^N$ denotes the empirical measure $\widehat{\rho}_{Y,t}^N := \frac{1}{N} \sum_{i=1}^N \delta_{Y_t^i}$ of the particles' local best positions. The choice of the weight $\omega_\alpha^\mathcal{E}$ in (1.4) comes from the well-known Laplace principle [12, 38], a classical asymptotic argument for integrals stating that for any probability measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ it holds

$$\lim_{\alpha \rightarrow \infty} \left(-\frac{1}{\alpha} \log \left(\int_{\mathbb{R}^d} \omega_\alpha^\mathcal{E}(y) d\varrho(y) \right) \right) = \inf_{y \in \text{supp}(\varrho)} \mathcal{E}(y). \quad (1.5)$$

Based thereon, $y_\alpha(\widehat{\rho}_{Y,t}^N)$ is expected to be a rough estimate for a global minimizer x^* , which improves as $\alpha \rightarrow \infty$ and as larger regions of the domain are explored. To feature the latter, the two remaining terms in (1.2c), each associated with a drift term, are diffusion terms injecting randomness into the dynamics through independent standard Brownian motions $((B_t^{1,i})_{t \geq 0})_{i=1, \dots, N}$ and $((B_t^{2,i})_{t \geq 0})_{i=1, \dots, N}$. The two commonly studied diffusion types for similar methods are isotropic [8, 18, 42] and anisotropic [9, 19] diffusion with

$$D(y - x) = \begin{cases} \|y - x\|_2 \text{Id}, & \text{for isotropic diffusion,} \\ \text{diag}(y - x), & \text{for anisotropic diffusion,} \end{cases} \quad (1.6)$$

where $\text{Id} \in \mathbb{R}^{d \times d}$ is the identity matrix and $\text{diag} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ the operator mapping a vector onto a diagonal matrix with the vector as its diagonal. Intuitively, the term's scaling encourages agents far from its own local best position or the globally computed consensus point to explore larger regions, whereas agents already close try to enhance their position only locally. As the coordinate-dependent scaling of anisotropic diffusion has been proven to be highly beneficial for high-dimensional problems [9, 17], in what follows, we limit our analysis to this case. An illustration of the formerly described PSO dynamics (1.2) is given in Fig. 1.

A theoretical convergence analysis of PSO is possible either on the microscopic level (1.2) or by analyzing the macroscopic behavior of the particle density through a mean-field limit, what usually admits more powerful analysis tools. In the large particle limit an individual particle is not influenced any more by individual particles but only by the average behavior of all particles. As shown in [21, Section 3.2], the empirical particle measure $\widehat{f}^N := \frac{1}{N} \sum_{i=1}^N \delta_{(X^i, Y^i, V^i)}$ converges in law to the deterministic agent distribution $f \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^{3d}))$, which weakly satisfies the nonlinear Vlasov-Fokker-Planck equation

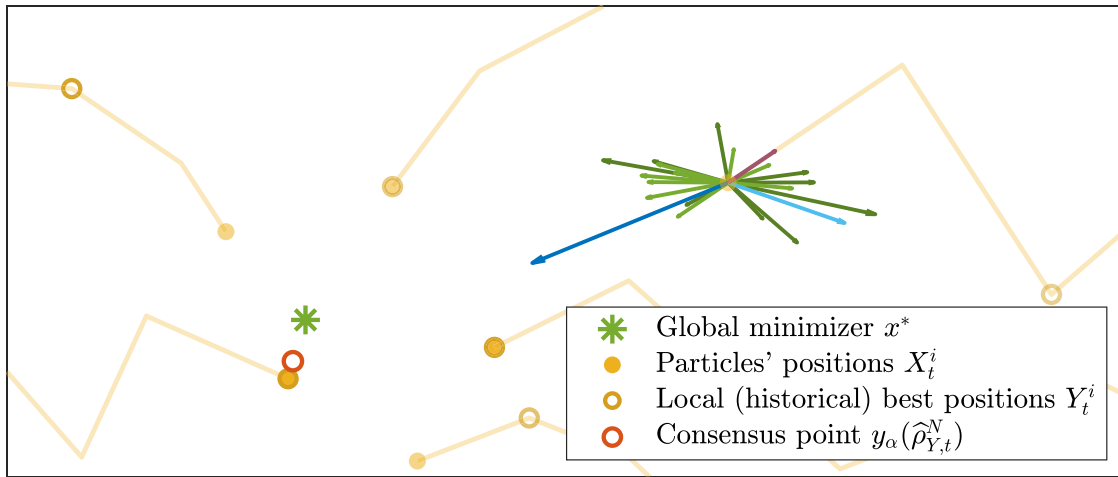


Fig. 1 An illustration of the PSO dynamics. Agents with positions X^1, \dots, X^N (yellow dots with their trajectories) explore the energy landscape of \mathcal{E} in search of the global minimizer x^* (green star). The dynamics of each particle is governed by five terms. A local drift term (light blue arrow) imposes a force towards its local best position Y_t^i (indicated by a circle). A global drift term (dark blue arrow) drags the agent towards a momentaneous consensus point $y_\alpha(\hat{\rho}_{Y,t}^N)$ (orange circle) computed as a weighted (visualized through color opacity) average of the particles' local best positions. Friction (purple arrow) counteracts inertia. The two remaining terms are diffusion terms (light and dark green arrows) associated with a respective drift term

$$\begin{aligned} & \partial_t f_t + v \cdot \nabla_x f_t + \nabla_y \cdot (\kappa(x - y) S^{\beta, \theta}(x, y) f_t) \\ &= \nabla_v \cdot \left(\frac{\gamma}{m} v f_t + \frac{\lambda_1}{m} (x - y) f_t + \frac{\lambda_2}{m} (x - y_\alpha(\rho_{Y,t})) f_t \right) \\ &+ \left(\frac{\sigma_1^2}{2m^2} (D(x - y))^2 + \frac{\sigma_2^2}{2m^2} (D(x - y_\alpha(\rho_{Y,t})))^2 \right) \nabla_v f_t \end{aligned} \quad (1.7)$$

with initial datum f_0 . The mean-field limit results [6, 25–27, 52] ensure that the particle system (1.2) is well-approximated by the following self-consistent mean-field McKean process

$$d\bar{X}_t = \bar{V}_t dt, \tag{1.8a}$$

$$d\bar{Y}_t = \kappa (\bar{X}_t - \bar{Y}_t) S^{\beta, \theta}(\bar{X}_t, \bar{Y}_t) dt, \tag{1.8b}$$

$$\begin{aligned} m d\bar{V}_t = & -\gamma \bar{V}_t dt + \lambda_1 (\bar{Y}_t - \bar{X}_t) dt + \lambda_2 (y_\alpha(\rho_{Y,t}) - \bar{X}_t) dt \\ & + \sigma_1 D(\bar{Y}_t - \bar{X}_t) dB_t^1 + \sigma_2 D(y_\alpha(\rho_{Y,t}) - \bar{X}_t) dB_t^2, \end{aligned} \tag{1.8c}$$

with initial datum $(\bar{X}_0, \bar{Y}_0, \bar{V}_0) \sim f_0$ and the marginal law $\rho_{Y,t}$ of \bar{Y}_t given by

$$\rho_Y(t, \cdot) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} df_t(x, \cdot, v).$$

Here, f_t denotes the distribution of $(\bar{X}_t, \bar{Y}_t, \bar{V}_t)$. This makes (1.7) and (1.8) nonlinear.

1.1 Contribution

In view of the versatility, efficiency, and wide applicability of PSO combined with its long historical tradition, a mathematical analysis of the finite particle system (1.2) is of considerable interest.

In this work we advance the theoretical understanding of the method and contribute to the completion of a full numerical analysis of PSO by proving rigorously the convergence of PSO with memory effects to global minimizers using mean-field techniques. More precisely, under mild regularity assumptions on the objective \mathcal{E} and a well-preparation condition about the initialization f_0 , we analyze the behavior of the particle distribution f solving the mean-field dynamics (1.8). At first, it is shown that concentration is achieved at some \tilde{x} in the sense that the marginal law w.r.t. the local best position, $\rho_{Y,t}$, converges narrowly to a Dirac delta $\delta_{\tilde{x}}$ as $t \rightarrow \infty$. Consecutively, we argue that, for an appropriate choice of the parameters, in particular $\alpha \gg 1$, which may depend on the dimension d , $\mathcal{E}(\tilde{x})$ can be made arbitrarily close to the minimal value $\underline{\mathcal{E}} := \inf_{x \in \mathbb{R}^d} \mathcal{E}(x)$. A suitable tractability condition on the objective \mathcal{E} eventually ensures that \tilde{x} is close to a global minimizer. Similar mean-field convergence results are obtained for the case without memory effects. In this setting we are moreover able to establish the convergence of the interacting N -particle dynamics to its mean-field limit with a dimension-independent rate, which allows to obtain a so far unique holistic and quantitative convergence statement of PSO. As the mean-field approximation result does not suffer from the curse of dimensionality, we in particular prove that the numerical PSO method has polynomial complexity. With these new results we solve the theoretical open problem about the convergence of PSO posed in [22].

Furthermore, we propose an efficient and parallelizable implementation, which is particularly suited for machine learning problems by integrating modern machine learning techniques such as random mini-batch ideas as well as traditional metaheuristic-inspired techniques from genetic programming and simulated annealing.

1.2 Prior Arts

The convergence of PSO algorithms has been investigated by many scholars since its introduction, which has lead to several variations allowing to establish desirable properties such as consensus formation or convergence to optimal solutions. While the matter of consensus is well-studied, see, e.g., [11, 40] or more recently [56], where the authors employ stochastic approximation methods [32], only few general theoretical statements regarding the properties of the found consensus are available. Both the existence of a large number of variations of the algorithm and the lack of a rigorous global convergence analysis are attributed amongst other things, such as the stochasticity and the usage of memory mechanisms, to the phenomenon of premature convergence of basic PSO [31], which was observed in [4, 5] and remedied by proposing a modified version, called guaranteed convergence PSO. Nevertheless, this adaptation only allows to prove the convergence to local optima. In order to obtain therefrom a stochastic global search algorithm, the authors suggest to add

purely stochastic particles to the swarm, which trivially makes the method capable of detecting a global optimizer, but entails a computational time which coincides with the time required to examine every location in the search space. Other works consider certain notions of weak convergence [7] or provide probabilistic guarantees of finding locally optimal solutions, meaning that eventually all particles are located almost surely at a local optimum of the objective function [51]. In [44], similarly to our work, the expected behavior of the particles is investigated.

All of the formerly mentioned results though are obtained through the analysis of the particles' trajectories generated by a time-discretized algorithm as in [21, Equation (6.3)]. The present paper takes a different point of view by studying the continuous-time description of the PSO model (1.2) through the lens of the mean-field approximation (1.7). Analyzing the macroscopic behavior of a system through a mean-field limit instead of investigating the microscopic particle dynamics has its origins in statistical mechanics [29], where interactions between particles are approximated by an averaged influence. By eliminating the correlation between the particles, a many-body problem can be reduced to a one-body problem, which is usually much easier to solve while still giving an understanding of the mechanisms at play by describing the average behavior of the particles. These ideas, for instance, are also used to study the collective behavior of animals when forming large-scale patterns through self-organization by analyzing an associated kinetic PDE [6]. In very recent works, this perspective of analysis has also been taken to demystify the training process of neural networks, see, e.g., [13, 36], where a mean-field approximation is utilized to formulate risk minimization by stochastic gradient descent (SGD) in terms of a gradient-flow PDE, which allows for a rigorous mathematical analysis.

The analysis technique we use follows the line of work of self-organization. It is inspired by [8, 9], where a variance-based analysis approach has been developed for consensus-based optimization (CBO), which follows the guiding principles of metaheuristics and in particular resembles PSO but is of much simpler nature and therefore easier to analyze. In comparison to Equation (1.2), CBO methods are described by a system of first-order SDEs [8, Equation (1.1)] and do not contain memory mechanisms, which are responsible for both a significantly more challenging mathematical modeling and convergence analysis.

1.3 Organization

Sections 2 and 3 are dedicated to the analysis of PSO without and with memory mechanisms, respectively. After providing details about the well-posedness of the mean-field dynamics, we present and discuss the main result about the convergence of the mean-field dynamics to a global minimizer of the objective function. In Sect. 4 we then state a quantitative result about the mean-field approximation for PSO without memory effects, which enables us to obtain a holistic convergence statement of the numerical PSO method. Eventually, a computationally efficient implementation of PSO is proposed in Sect. 5, before Sect. 6 concludes the paper. For the sake of reproducible research, in the GitHub repository <https://github.com/KonstantinRiedl/PSOAnalysis> we provide the Matlab code implementing the PSO algorithm analyzed in this work.

2 Mean-Field Analysis of PSO Without Memory Effects

Before providing a theoretical analysis of the mean-field PSO dynamics (1.7) and (1.8), in this section we investigate a reduced version, which does not involve memory mechanisms. Its multi-particle formulation was proposed in [22, Section 3.1] and reads

$$dX_t^i = V_t^i dt, \quad (2.1a)$$

$$m dV_t^i = -\gamma V_t^i dt + \lambda \left(x_\alpha(\widehat{\rho}_{X,t}^N) - X_t^i \right) dt + \sigma D \left(x_\alpha(\widehat{\rho}_{X,t}^N) - X_t^i \right) dB_t^i. \quad (2.1b)$$

Compared to the full model, each particle is characterized only by its position X^i and velocity V^i . The forces acting on a particle, i.e., influencing its velocity in Equation (2.1b), are friction, acceleration through the consensus drift and diffusion as in (1.6) with independent standard Brownian motions $((B_t^i)_{t \geq 0})_{i=1, \dots, N}$. The consensus point $x_\alpha(\widehat{\rho}_{X,t}^N)$ is directly computed from the current positions of the particles according to

$$x_\alpha(\widehat{\rho}_{X,t}^N) := \int_{\mathbb{R}^d} x \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L^1(\widehat{\rho}_{X,t}^N)}} d\widehat{\rho}_{X,t}^N(x), \quad (2.2)$$

where $\widehat{\rho}_{X,t}^N$ denotes the empirical measure $\widehat{\rho}_{X,t}^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$ of the particles' positions. Independent and identically distributed initial data $((X_0^i, V_0^i) \sim f_0)_{i=1, \dots, N}$ with $f_0 \in \mathcal{P}(\mathbb{R}^{2d})$ complement (2.1).

Similar to the particle system (1.2), as $N \rightarrow \infty$, the mean-field dynamics of (2.1) is described by the nonlinear self-consistent McKean process

$$d\bar{X}_t = \bar{V}_t dt, \quad (2.3a)$$

$$m d\bar{V}_t = -\gamma \bar{V}_t dt + \lambda \left(x_\alpha(\rho_{X,t}) - \bar{X}_t \right) dt + \sigma D \left(x_\alpha(\rho_{X,t}) - \bar{X}_t \right) dB_t, \quad (2.3b)$$

with initial datum $(\bar{X}_0, \bar{V}_0) \sim f_0$ and the marginal law $\rho_{X,t}$ of \bar{X}_t given by $\rho_X(t, \cdot) = \int_{\mathbb{R}^d} df(t, \cdot, v)$. A direct application of the Itô-Doebelin formula shows that the law $f \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^{2d}))$ is a weak solution to the nonlinear Vlasov-Fokker-Planck equation

$$\begin{aligned} & \partial_t f_t + v \cdot \nabla_x f_t \\ & = \nabla_v \cdot \left(\frac{\gamma}{m} v f_t + \frac{\lambda}{m} (x - x_\alpha(\rho_{X,t})) f_t + \frac{\sigma^2}{2m^2} (D(x - x_\alpha(\rho_{X,t})))^2 \nabla_v f_t \right) \end{aligned} \quad (2.4)$$

with initial datum f_0 .

Remark 2 A separate theoretical analysis of the dynamics (2.1) is necessary as it cannot be derived from (1.2) in a way that also the proof technique can be adopted in a straightforward manner. This can be seen from subtle differences in the proofs of Theorems 2 and 4; see in particular Lemma 3.

It is also worth noting that Equation (2.1) bears a certain resemblance to CBO [8, 9, 18, 19, 42], whereas (1.8) resembles [48]. Indeed, as made rigorous in [10], CBO methods can be derived from PSO in the small inertia limit $m \rightarrow 0$, or equivalently $\gamma \rightarrow 1$. Nevertheless, analyzing the convergence of CBO directly permits sharper bounds when compared to utilizing the results obtained in our work together with [10, Theorem 2.4].

Before turning towards the well-posedness of the mean-field dynamics (2.3) and presenting the main result of this section about the convergence to the global minimizer x^* , let us introduce the class of objective function \mathcal{E} considered in the theoretical part of this work. We remark that the assumptions made in what follows coincide with the ones of [8, 9] as well as several subsequent works in this direction.

Assumption 1 Throughout the paper we are interested in objective functions $\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}$, for which

- A1 there exists $x^* \in \mathbb{R}^d$ such that $\mathcal{E}(x^*) = \inf_{x \in \mathbb{R}^d} \mathcal{E}(x) =: \underline{\mathcal{E}}$,
- A2 there exists some constant $L_{\mathcal{E}} > 0$ such that

$$|\mathcal{E}(x) - \mathcal{E}(x')| \leq L_{\mathcal{E}} (|x| + |x'|) |x - x'|, \quad \text{for all } x, x' \in \mathbb{R}^d,$$

- A3 either $\bar{\mathcal{E}} := \sup_{x \in \mathbb{R}^d} \mathcal{E}(x) < \infty$ or there exist constants $c_{\mathcal{E}}, R > 0$ such that

$$\mathcal{E}(x) - \underline{\mathcal{E}} \geq c_{\mathcal{E}} |x|^2, \quad \text{for all } x \in \mathbb{R}^d \text{ with } |x| \geq R,$$

- A4 $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ with $\|\nabla^2 \mathcal{E}\|_{\infty} \leq C_{\mathcal{E}}$ for some constant $C_{\mathcal{E}} > 0$,
- A5 there exist $\eta > 0$ and $\nu \in (0, \infty)$ such that for any $x \in \mathbb{R}^d$ there exists a global minimizer x^* of \mathcal{E} (which may depend on x) such that

$$|x - x^*| \leq (\mathcal{E}(x) - \underline{\mathcal{E}})^{\nu} / \eta.$$

Assumption A1 just states that the objective function \mathcal{E} attains its infimum $\underline{\mathcal{E}}$ at some $x^* \in \mathbb{R}^d$, which may not necessarily be unique. Assumption A2 describes the local Lipschitz-continuity of \mathcal{E} , entailing in particular that the objective has at most quadratic growth at infinity. Assumption A3, on the other hand, requires \mathcal{E} to be either bounded or of at least quadratic growth in the farfield. Together, A2 and A3 allow to obtain the well-posedness of the PSO model. Assumption A4 is a regularity assumption about \mathcal{E} , which is required only for the theoretical analysis. The quadratic growth nature of Assumptions A2–A4 in the farfield may bear a certain resemblance to log-Sobolev inequalities [50], which are pivotal in the convergence analysis of simulated annealing, see [53] for further details. Unlike simulated annealing however, the PSO method is a zero-order method where we do not need the gradient information of the objective function in the numerical application. Assumption A5 should be interpreted as a tractability condition of the landscape of \mathcal{E} , which ensures that achieving an objective value of approximately $\underline{\mathcal{E}}$ guarantees closeness to a global minimizer x^* and thus eliminates cases of almost-optimal valleys in the energy landscape far away from

any globally minimizing argument. Such assumption is therefore also referred to as an inverse continuity property.

It shall be emphasized that objectives with multiple global minima of identical quality are not excluded.

2.1 Well-Posedness of PSO without Memory Effects

Let us recite a well-posedness result about the mean-field PSO dynamics (2.3) and the associated Vlasov-Fokker-Planck equation (2.4). Its proof is analogous to the one provided for Theorem 3 for the full dynamics (1.8) based on the Leray-Schauder fixed point theorem.

Theorem 1 *Let \mathcal{E} satisfy Assumptions A1–A3. Moreover, let $m, \gamma, \lambda, \sigma, \alpha, T > 0$. If (\bar{X}_0, \bar{V}_0) is distributed according to $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$, then the nonlinear SDE (2.3) admits a unique strong solution up to time T with the paths of process (\bar{X}, \bar{V}) valued in $\mathcal{C}([0, T], \mathbb{R}^d) \times \mathcal{C}([0, T], \mathbb{R}^d)$. The associated law f has regularity $\mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ and is a weak solution to the Vlasov-Fokker-Planck equation (2.4). In particular,*

$$\sup_{t \in [0, T]} \mathbb{E}[|\bar{X}_t|^4 + |\bar{V}_t|^4] \leq \left(1 + 2\mathbb{E}[|\bar{X}_0|^4 + |\bar{V}_0|^4]\right) e^{CT} \quad (2.5)$$

for some constant $C > 0$ depending only on $m, \gamma, \lambda, \sigma, \alpha, c_{\mathcal{E}}, R$, and $L_{\mathcal{E}}$.

2.2 Convergence of PSO without Memory Effects to a Global Minimizer

A successful application of the PSO dynamics underlies the premise that the particles form consensus about a certain position \tilde{x} . In particular, in the mean-field limit one expects that the distribution of a particle's position $\rho_{X,t}$ converges to a Dirac delta $\delta_{\tilde{x}}$. This entails that the variance in the position $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]$ and the second-order moment of the velocity $\mathbb{E}[|\bar{V}_t|^2]$ of the averaged particle vanish. As we show in what follows, both functionals indeed decay exponentially fast in time. Motivated by these expectations we define the functional

$$\mathcal{H}(t) := \left(\frac{\gamma}{2m}\right)^2 |\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2 + \frac{\gamma}{2m} \langle \bar{X}_t - \mathbb{E}[\bar{X}_t], \bar{V}_t \rangle, \quad (2.6)$$

which we analyze in the remainder of this section. Its last term is required from a technical perspective. However, by proving the decay of $\mathbb{E}[\mathcal{H}(t)]$, which acts as Lyapunov function of the dynamics, one immediately obtains the same for $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2]$ as a consequence of the equivalence established in Lemma 1, which follows from Young's inequality.

Lemma 1 *The functional $\mathcal{H}(t)$ is equivalent to $|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2$ in the sense that*

$$\begin{aligned} & \frac{1}{2} \left(\frac{\gamma}{2m}\right)^2 |\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + \frac{1}{2} |\bar{V}_t|^2 \\ & \leq \mathcal{H}(t) \leq \frac{3}{2} \left(\left(\frac{\gamma}{2m}\right)^2 + 1 \right) \left(|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2 \right). \end{aligned} \tag{2.7}$$

We now derive an evolution inequality of the Lyapunov function $\mathbb{E}[\mathcal{H}(t)]$.

Lemma 2 *Let \mathcal{E} satisfy Assumptions A1–A3 and let $(\bar{X}_t, \bar{V}_t)_{t \geq 0}$ be a solution to the nonlinear SDE (2.3). Then $\mathbb{E}[\mathcal{H}(t)]$ with \mathcal{H} as defined in (2.6) satisfies*

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] & \leq -\frac{\gamma}{m} \mathbb{E}[|\bar{V}_t|^2] \\ & - \left(\frac{\lambda\gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2} \right) \frac{2e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))]} \right) \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]. \end{aligned} \tag{2.8}$$

Proof Let us write $\delta\bar{X}_t := \bar{X}_t - \mathbb{E}[\bar{X}_t]$ for short and note that the integration by parts formula gives

$$\frac{d}{dt} \mathbb{E}[|\delta\bar{X}_t|^2] = 2\mathbb{E}[\langle \delta\bar{X}_t, \bar{V}_t \rangle]. \tag{2.9}$$

Observe that, in what follows, the appearing stochastic integrals have vanishing expectations as a consequence of the regularity $f \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ obtained in Theorem 1. This is due to [39, Theorem 3.2.1(iii), Definition 3.1.4(iii)], which state that a stochastic integral vanishes if the associated second moment is integrable. Notice that the latter condition is sufficient for the stochastic integral to be a martingale. Applying the Itô-Doebelin formula and Young’s inequality yields

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[|\bar{V}_t|^2] & = -\frac{2\gamma}{m} \mathbb{E}[|\bar{V}_t|^2] + \frac{2\lambda}{m} \mathbb{E}[\langle \bar{V}_t, x_\alpha(\rho_{X,t}) - \bar{X}_t \rangle] + \frac{\sigma^2}{m^2} \mathbb{E}[|x_\alpha(\rho_{X,t}) - \bar{X}_t|^2] \\ & \leq -\left(\frac{2\gamma}{m} - \frac{\lambda}{\varepsilon m}\right) \mathbb{E}[|\bar{V}_t|^2] + \left(\frac{\varepsilon\lambda}{m} + \frac{\sigma^2}{m^2}\right) \mathbb{E}[|x_\alpha(\rho_{X,t}) - \bar{X}_t|^2], \quad \forall \varepsilon > 0. \end{aligned} \tag{2.10}$$

Again by employing the Itô-Doebelin formula we obtain

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\langle \delta\bar{X}_t, \bar{V}_t \rangle] & = \mathbb{E}[|\bar{V}_t|^2] - (\mathbb{E}[\bar{V}_t])^2 - \frac{\gamma}{m} \mathbb{E}[\langle \delta\bar{X}_t, \bar{V}_t \rangle] + \frac{\lambda}{m} \mathbb{E}[\langle \delta\bar{X}_t, x_\alpha(\rho_{X,t}) - \bar{X}_t \rangle] \\ & \leq \mathbb{E}[|\bar{V}_t|^2] - \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta\bar{X}_t|^2] + \frac{\lambda}{m} \mathbb{E}[\langle \delta\bar{X}_t, x_\alpha(\rho_{X,t}) - \mathbb{E}[\bar{X}_t] \rangle] - \frac{\lambda}{m} \mathbb{E}[|\delta\bar{X}_t|^2] \\ & = \mathbb{E}[|\bar{V}_t|^2] - \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta\bar{X}_t|^2] - \frac{\lambda}{m} \mathbb{E}[|\delta\bar{X}_t|^2], \end{aligned} \tag{2.11}$$

where we used the identity (2.9) and the fact that $\mathbb{E}[\langle \delta \bar{X}_t, x_\alpha(\rho_{X,t}) - \mathbb{E}[\bar{X}_t] \rangle] = 0$ in the last two steps. We now rearrange inequality (2.11) to get

$$\frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta \bar{X}_t|^2] + \frac{d}{dt} \mathbb{E}[\langle \delta \bar{X}_t, \bar{V}_t \rangle] \leq \mathbb{E}[|\bar{V}_t|^2] - \frac{\lambda}{m} \mathbb{E}[|\delta \bar{X}_t|^2],$$

which, in combination with (2.10), allows to show

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] &\leq - \left(\frac{3\gamma}{2m} - \frac{\lambda}{\varepsilon m} \right) \mathbb{E}[|\bar{V}_t|^2] - \frac{\lambda\gamma}{2m^2} \mathbb{E}[|\delta \bar{X}_t|^2] \\ &\quad + \left(\frac{\varepsilon\lambda}{m} + \frac{\sigma^2}{m^2} \right) \mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2]. \end{aligned} \tag{2.12}$$

In order to upper bound $\mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2]$, an application of Jensen’s inequality yields

$$\begin{aligned} \mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2] &\leq \frac{\iint_{\mathbb{R}^{2d}} |x - x'|^2 \omega_\alpha^\mathcal{E}(x') d\rho_{X,t}(x') d\rho_{X,t}(x)}{\int_{\mathbb{R}^d} \omega_\alpha^\mathcal{E}(x') d\rho_{X,t}(x')} \\ &\leq 2e^{-\alpha\mathcal{E}} \frac{\mathbb{E}[|\delta \bar{X}_t|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))]} \end{aligned} \tag{2.13}$$

By choosing $\varepsilon = (2\lambda)/\gamma$ in (2.12) and utilizing the estimate (2.13), we obtain (2.8) as desired. □

Remark 3 To obtain exponential decay of $\mathbb{E}[\mathcal{H}(t)]$ it is necessary to ensure the negativity of the prefactor of $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]$ in Inequality (2.8) by choosing the parameters of the PSO method in a suitable manner. This may be achieved by choosing for any fixed time t , given α and arbitrary $\sigma, \gamma > 0$,

$$\lambda > 4D_t^X \sigma^2 / \gamma \quad \text{and subsequently} \quad m < \gamma^2 / (8D_t^X \lambda), \tag{2.14}$$

where we abbreviate $D_t^X = 2e^{-\alpha\mathcal{E}} / \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))]$.

In order to be able to choose the parameters in Remark 3 once at the beginning of the algorithm instead of at every time step t , we need to be able to control the time-evolution of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))]$. We therefore study its time-derivative in the following lemma.

Lemma 3 *Let \mathcal{E} satisfy Assumptions A1–A4 and let $(\bar{X}_t, \bar{V}_t)_{t \geq 0}$ be the solution to the nonlinear SDE (2.3). Then it holds that*

$$\begin{aligned} \frac{d^2}{dt^2} (\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))])^2 &\geq -\frac{\gamma}{m} \frac{d}{dt} (\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))])^2 \\ &\quad - 4\alpha e^{-2\alpha\mathcal{E}} C_\mathcal{E} \left(1 + 2\frac{\lambda}{m} \left(\frac{2m}{\gamma} \right)^2 \right) \mathbb{E}[\mathcal{H}(t)]. \end{aligned} \tag{2.15}$$

Proof We first note that

$$\begin{aligned}
 & \frac{1}{2} \frac{d^2}{dt^2} (\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))]^2 \\
 &= \frac{d}{dt} \left(\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \frac{d}{dt} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \right) \\
 &= \left(\frac{d}{dt} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \right)^2 \\
 &\quad + \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \frac{d^2}{dt^2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \\
 &\geq \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \frac{d^2}{dt^2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))], \tag{2.16}
 \end{aligned}$$

leaving the second time-derivative of $\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))]$ to be lower bounded. To do so, we start with its first derivative. Applying the Itô-Doeblin formula twice and noting that stochastic integrals have vanishing expectations as a consequence of [39, Theorem 3.2.1(iii), Definition 3.1.4(iii)] combined with the regularity $f \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ obtained in Theorem 1, we have

$$\begin{aligned}
 \frac{d}{dt} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] &= -\alpha \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t)) \langle \nabla \mathcal{E}(\bar{X}_t), \bar{V}_t \rangle] \\
 &= -\alpha \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0)) \langle \nabla \mathcal{E}(\bar{X}_0), \bar{V}_0 \rangle] \\
 &\quad - \alpha \mathbb{E} \left[\int_0^t d \langle \exp(-\alpha \mathcal{E}(\bar{X}_s)) \nabla \mathcal{E}(\bar{X}_s), \bar{V}_s \rangle \right] \\
 &= -\alpha \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0)) \langle \nabla \mathcal{E}(\bar{X}_0), \bar{V}_0 \rangle] \\
 &\quad - \alpha \mathbb{E} \left[\int_0^t \langle \exp(-\alpha \mathcal{E}(\bar{X}_s)) \bar{V}_s, \nabla^2 \mathcal{E}(\bar{X}_s) \bar{V}_s \rangle ds \right] \tag{2.17} \\
 &\quad + \alpha^2 \mathbb{E} \left[\int_0^t \exp(-\alpha \mathcal{E}(\bar{X}_s)) |\langle \nabla \mathcal{E}(\bar{X}_s), \bar{V}_s \rangle|^2 ds \right] \\
 &\quad - \alpha \mathbb{E} \left[\int_0^t \exp(-\alpha \mathcal{E}(\bar{X}_s)) \left\langle \nabla \mathcal{E}(\bar{X}_s), -\frac{\gamma}{m} \bar{V}_s \right\rangle ds \right] \\
 &\quad - \alpha \mathbb{E} \left[\int_0^t \exp(-\alpha \mathcal{E}(\bar{X}_s)) \left\langle \nabla \mathcal{E}(\bar{X}_s), \frac{\lambda}{m} (x_\alpha(\rho_{X,s}) - \bar{X}_s) \right\rangle ds \right].
 \end{aligned}$$

Differentiating both sides of (2.17) with respect to the time t yields

$$\begin{aligned}
 \frac{d^2}{dt^2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] &= -\alpha \mathbb{E}[\langle \exp(-\alpha \mathcal{E}(\bar{X}_t)) \bar{V}_t, \nabla^2 \mathcal{E}(\bar{X}_t) \bar{V}_t \rangle] \\
 &\quad + \alpha^2 \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t)) |\langle \nabla \mathcal{E}(\bar{X}_t), \bar{V}_t \rangle|^2] \\
 &\quad + \frac{\alpha \gamma}{m} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t)) \langle \nabla \mathcal{E}(\bar{X}_t), \bar{V}_t \rangle]
 \end{aligned}$$

$$\begin{aligned}
 & - \frac{\alpha\lambda}{m} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t)) \langle \nabla \mathcal{E}(\bar{X}_t), x_\alpha(\rho_{X,t}) - \bar{X}_t \rangle] \\
 \geq & - \frac{\gamma}{m} \frac{d}{dt} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))] \\
 & - \underbrace{\alpha \mathbb{E}[\langle \exp(-\alpha\mathcal{E}(\bar{X}_t)) \bar{V}_t, \nabla^2 \mathcal{E}(\bar{X}_t) \bar{V}_t \rangle]}_{T_1} \\
 & - \frac{\alpha\lambda}{m} \underbrace{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t)) \langle \nabla \mathcal{E}(\bar{X}_t), x_\alpha(\rho_{X,t}) - \bar{X}_t \rangle]}_{T_2},
 \end{aligned} \tag{2.18}$$

where we employed the first line of (2.17) in the last step. It remains to upper bound the terms T_1 and T_2 . Making use of Assumptions A1 and A4, we immediately obtain

$$T_1 \leq \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t)) \|\nabla^2 \mathcal{E}\|_\infty |\bar{V}_t|^2] \leq e^{-\alpha\underline{\mathcal{E}}} C_{\mathcal{E}} \mathbb{E}[|\bar{V}_t|^2]. \tag{2.19}$$

For T_2 , again under Assumptions A1 and A4, we first note that

$$\begin{aligned}
 T_2 & = -\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t)) \langle \nabla \mathcal{E}(\bar{X}_t) - \nabla \mathcal{E}(x_\alpha(\rho_{X,t})), \bar{X}_t - x_\alpha(\rho_{X,t}) \rangle] \\
 & \leq e^{-\alpha\underline{\mathcal{E}}} C_{\mathcal{E}} \mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2],
 \end{aligned} \tag{2.20}$$

where the equality is a consequence of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t)) \langle \nabla \mathcal{E}(x_\alpha(\rho_{X,t})), \bar{X}_t - x_\alpha(\rho_{X,t}) \rangle] = 0$, which follows from the definition of $x_\alpha(\rho_{X,t})$. Bounding $\mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2]$ as in (2.13) we can further bound (2.20) as

$$T_2 \leq e^{-\alpha\underline{\mathcal{E}}} C_{\mathcal{E}} \mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2] \leq 2e^{-2\alpha\underline{\mathcal{E}}} C_{\mathcal{E}} \frac{\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))]} \tag{2.21}$$

Collecting the estimates (2.19) and (2.21) within (2.18) and inserting the result into (2.16) give

$$\begin{aligned}
 \frac{1}{2} \frac{d^2}{dt^2} (\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))])^2 & \geq - \frac{\gamma}{m} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))] \frac{d}{dt} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))] \\
 & \quad - \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))] \alpha C_{\mathcal{E}} e^{-\alpha\underline{\mathcal{E}}} \mathbb{E}[|\bar{V}_t|^2] \\
 & \quad - \frac{2\alpha\lambda}{m} e^{-2\alpha\underline{\mathcal{E}}} C_{\mathcal{E}} \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] \\
 & \geq - \frac{\gamma}{2m} \frac{d}{dt} (\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))])^2 \\
 & \quad - \alpha e^{-2\alpha\underline{\mathcal{E}}} C_{\mathcal{E}} \left(\mathbb{E}[|\bar{V}_t|^2] + \frac{2\lambda}{m} \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] \right),
 \end{aligned}$$

which yields the statement after employing the lower bound of (2.7) as in Lemma 1. □

We are now ready to state and prove the main result about the convergence of the mean-field PSO dynamics (2.3) without memory mechanisms to the global minimizer x^* .

Theorem 2 *Let \mathcal{E} satisfy Assumptions A1–A4 and let $(\bar{X}_t, \bar{V}_t)_{t \geq 0}$ be a solution to the nonlinear SDE (2.3). Moreover, let us assume the well-preparation of the initial datum \bar{X}_0 and \bar{V}_0 in the sense that*

P1 $\mu > 0$ with

$$\mu := \frac{\lambda\gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2} \right) \frac{4e^{-\alpha\underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_0))]},$$

P2 it holds

$$\begin{aligned} & \frac{m\alpha}{2\gamma} \frac{(\mathbb{E}[\langle \exp(-\alpha\mathcal{E}(\bar{X}_0)) \nabla \mathcal{E}(\bar{X}_0), \bar{V}_0 \rangle])_+}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_0))]} \\ & + \frac{\alpha C_{\mathcal{E}}}{\chi \left(\frac{\gamma}{m} - \chi \right)} \left(1 + \frac{8m\lambda}{\gamma^2} \right) \frac{\mathbb{E}[\mathcal{H}(0)]}{(\mathbb{E}[\exp(-\alpha(\mathcal{E}(\bar{X}_0) - \underline{\mathcal{E}}))])^2} < \frac{3}{16}, \end{aligned}$$

with $x_+ = \max\{x, 0\}$ for $x \in \mathbb{R}$ denoting the positive part and where

$$\chi := \frac{2}{3} \frac{\min\{\gamma/m, \mu\}}{((\gamma/(2m))^2 + 1)}.$$

Then $\mathbb{E}[\mathcal{H}(t)]$ with \mathcal{H} as defined in Equation (2.6) converges exponentially fast with rate χ to 0 as $t \rightarrow \infty$. Moreover, there exists some \tilde{x} , which may depend on α and f_0 , such that $\mathbb{E}[\bar{X}_t] \rightarrow \tilde{x}$ and $x_\alpha(\rho_{X,t}) \rightarrow \tilde{x}$ exponentially fast with rate $\chi/2$ as $t \rightarrow \infty$. Eventually, for any given accuracy $\varepsilon > 0$, there exists $\alpha_0 > 0$, which may depend on the dimension d , such that for all $\alpha > \alpha_0$, \tilde{x} satisfies

$$\mathcal{E}(\tilde{x}) - \underline{\mathcal{E}} \leq \varepsilon.$$

If \mathcal{E} additionally satisfies Assumption A5, we have $|\tilde{x} - x^| \leq \varepsilon^\nu/\eta$.*

Remark 4 As suggested in Remark 3, Theorem 2 traces back the evolution of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))]$ to its initial state by employing Lemma 3. This allows to fixate all parameters of PSO at initialization time. By replacing D_t^X with $2D_0^X$ in (2.14), the well-preparation of the parameters as in Condition P1 can be ensured.

Condition P2 requires the well-preparation of the initialization in the sense that the initial datum f_0 is both well-concentrated and to a certain extent not too far from an optimal value. While this might have a locality flavor, the condition is generally fulfilled in practical applications. Moreover, for CBO methods there is recent work where such assumption about the initial datum is reduced to the absolute minimum [18, 19].

Remark 5 The choice of the parameter α_0 necessary in Theorem 2 may be affected by the dimensionality d of the optimization problem at hand. By establishing a quantitative nonasymptotic Laplace principle, this dependence is made explicit in the works [18, Proposition 18] and [19, Proposition 1], where the authors show that α_0 may be required to grow linearly in d , see [18, Remark 21].

Proof of Theorem 2 Let us define the time horizon

$$T := \inf \left\{ t \geq 0 : \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] < \frac{1}{2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))] \right\} \quad \text{with } \inf \emptyset = \infty.$$

Obviously, by continuity, $T > 0$. We claim that $T = \infty$, which we prove by contradiction in the following. Therefore, assume $T < \infty$. Then, for $t \in [0, T]$, we have

$$\begin{aligned} & \frac{\lambda \gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2} \right) \frac{2e^{-\alpha \underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))]} \\ & \geq \frac{\lambda \gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2} \right) \frac{4e^{-\alpha \underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))]} = \mu > 0, \end{aligned}$$

where the positivity of μ is due to the well-preparation condition P1 of the initialization. Lemma 2 then provides an upper bound for the time derivative of the functional $\mathbb{E}[\mathcal{H}(t)]$,

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] & \leq -\frac{\gamma}{m} \mathbb{E}[|\bar{V}_t|^2] - \mu \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] \\ & \leq -\min \left\{ \frac{\gamma}{m}, \mu \right\} \left(\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] + \mathbb{E}[|\bar{V}_t|^2] \right) \quad (2.22) \\ & \leq -\frac{2}{3} \frac{\min\{\gamma/m, \mu\}}{((\gamma/(2m))^2 + 1)} \mathbb{E}[\mathcal{H}(t)] =: -\chi \mathbb{E}[\mathcal{H}(t)], \end{aligned}$$

where we made use of the upper bound of (2.7) as in Lemma 1 in the last inequality. The rate χ is defined implicitly and it is straightforward to check that $\chi < \gamma/m$. Grönwall’s inequality implies

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t). \quad (2.23)$$

Let us now investigate the evolution of the functional $\mathcal{X}(t) := (\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))])^2$. First note that

$$\dot{\mathcal{X}}(0) := \frac{d}{dt} \mathcal{X}(t)|_{t=0} = -2\alpha \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))] \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0)) \langle \nabla \mathcal{E}(\bar{X}_0), \bar{V}_0 \rangle].$$

Then, an application of Grönwall’s inequality to Equation (2.15) from Lemma 3 and using the explicit bound of $\mathbb{E}[\mathcal{H}(t)]$ from (2.23) yields

$$\begin{aligned} \frac{d}{dt} \mathcal{X}(t) &\geq \dot{\mathcal{X}}(0) \exp\left(-\frac{\gamma}{m}t\right) \\ &\quad - 4\alpha e^{-2\alpha \underline{\mathcal{E}}} C_{\mathcal{E}} \left(1 + 2\frac{\lambda}{m} \left(\frac{2m}{\gamma}\right)^2\right) \int_0^t \mathbb{E}[\mathcal{H}(s)] \exp\left(-\frac{\gamma}{m}(t-s)\right) ds \\ &\geq \dot{\mathcal{X}}(0) \exp\left(-\frac{\gamma}{m}t\right) \\ &\quad - 4\alpha e^{-2\alpha \underline{\mathcal{E}}} C_{\mathcal{E}} \left(1 + 2\frac{\lambda}{m} \left(\frac{2m}{\gamma}\right)^2\right) \mathbb{E}[\mathcal{H}(0)] \frac{1}{\gamma/m - \chi} \left(\exp(-\chi t) - \exp\left(-\frac{\gamma}{m}t\right)\right) \\ &\geq \dot{\mathcal{X}}(0) \exp\left(-\frac{\gamma}{m}t\right) \\ &\quad - 4\alpha e^{-2\alpha \underline{\mathcal{E}}} C_{\mathcal{E}} \left(1 + 2\frac{\lambda}{m} \left(\frac{2m}{\gamma}\right)^2\right) \mathbb{E}[\mathcal{H}(0)] \frac{1}{\gamma/m - \chi} \exp(-\chi t), \end{aligned}$$

which, in turn, implies

$$\mathcal{X}(t) \geq \mathcal{X}(0) - \frac{m}{\gamma} (-\dot{\mathcal{X}}(0))_+ - \frac{4\alpha e^{-2\alpha \underline{\mathcal{E}}} C_{\mathcal{E}}}{\chi(\gamma/m - \chi)} \left(1 + 2\frac{\lambda}{m} \left(\frac{2m}{\gamma}\right)^2\right) \mathbb{E}[\mathcal{H}(0)]$$

after discarding the positive parts. Recalling the definition of \mathcal{X} and employing the second well-preparation condition P2, we can deduce that for all $t \in [0, T]$ it holds

$$\begin{aligned} (\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))])^2 &\geq (\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))])^2 \\ &\quad - \frac{2m\alpha}{\gamma} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))] (\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0)) \langle \nabla \mathcal{E}(\bar{X}_0), \bar{V}_0 \rangle])_+ \\ &\quad - \frac{4\alpha e^{-2\alpha \underline{\mathcal{E}}} C_{\mathcal{E}}}{\chi(\gamma/m - \chi)} \left(1 + 2\frac{\lambda}{m} \left(\frac{2m}{\gamma}\right)^2\right) \mathbb{E}[\mathcal{H}(0)] \\ &> \frac{1}{4} (\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))])^2, \end{aligned}$$

which entails that there exists $\delta > 0$ such that $\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \geq \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))]/2$ in $[T, T + \delta]$ as well, contradicting the definition of T and therefore showing the claim $T = \infty$.

As a consequence of (2.23) we have

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t) \quad \text{and} \quad \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_t))] \geq \frac{1}{2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{X}_0))] \tag{2.24}$$

for all $t \geq 0$. In particular, by means of Lemma 1, for a suitable generic constant $C > 0$, we infer

$$\begin{aligned} \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] &\leq C \exp(-\chi t), \quad \mathbb{E}[|\bar{V}_t|^2] \leq C \exp(-\chi t), \\ \text{and } \mathbb{E}[|\bar{X}_t - x_\alpha(\rho_{X,t})|^2] &\leq C \exp(-\chi t), \end{aligned} \tag{2.25}$$

where the last inequality uses the fact (2.13). Moreover, with Jensen’s inequality,

$$\left| \frac{d}{dt} \mathbb{E}[\bar{X}_t] \right| \leq \mathbb{E}[|\bar{V}_t|] \leq C \exp(-\chi t/2) \rightarrow 0 \text{ as } t \rightarrow \infty,$$

showing that $\mathbb{E}[\bar{X}_t] \rightarrow \tilde{x}$ for some $\tilde{x} \in \mathbb{R}^d$, which may depend on α and f_0 . According to (2.25), $\bar{X}_t \rightarrow \tilde{x}$ in mean-square and $x_\alpha(\rho_{X,t}) \rightarrow \tilde{x}$, since

$$\begin{aligned} |x_\alpha(\rho_{X,t}) - \tilde{x}|^2 &\leq 3\mathbb{E}[|x_\alpha(\rho_{X,t}) - \bar{X}_t|^2] \\ &\quad + 3\mathbb{E}[|\bar{X}_t - \mathbb{E}\bar{X}_t|^2] + 3|\mathbb{E}\bar{X}_t - \tilde{x}|^2 \rightarrow 0 \text{ as } t \rightarrow \infty. \end{aligned}$$

Eventually, by continuity of the objective function \mathcal{E} and by the dominated convergence theorem, we conclude that $\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))] \rightarrow e^{-\alpha\mathcal{E}(\tilde{x})}$ as $t \rightarrow \infty$. Using this when taking the limit $t \rightarrow \infty$ in the second bound of (2.24) after applying the logarithm and multiplying both sides with $-1/\alpha$, we obtain

$$\mathcal{E}(\tilde{x}) = \lim_{t \rightarrow \infty} \left(-\frac{1}{\alpha} \log \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t))] \right) \leq -\frac{1}{\alpha} \log \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_0))] + \frac{1}{\alpha} \log 2. \tag{2.26}$$

The Laplace principle (1.5) on the other hand allows to choose $\tilde{\alpha} \gg 1$ large enough such that for given $\varepsilon > 0$ it holds $-\frac{1}{\tilde{\alpha}} \log \mathbb{E}[\exp(-\tilde{\alpha}\mathcal{E}(\bar{X}_0))] - \underline{\mathcal{E}} < \varepsilon/2$ for any $\alpha \geq \tilde{\alpha}$. Together with (2.26), this establishes $0 \leq \mathcal{E}(\tilde{x}) - \underline{\mathcal{E}} \leq \varepsilon/2 + (\log 2)/\alpha \leq \varepsilon$ for $\alpha \geq \max\{\tilde{\alpha}, (2 \log 2)/\varepsilon\}$. Finally, under the inverse continuity property A5 we additionally have $|\tilde{x} - x^*| \leq (\mathcal{E}(\tilde{x}) - \underline{\mathcal{E}})^\nu/\eta \leq \varepsilon^\nu/\eta$, concluding the proof. \square

3 Mean-Field Analysis of PSO with Memory Effects

Let us now turn back to the PSO dynamics (1.2) described in the introduction. The fundamental difference to what was analyzed in the preceding section is the presence of a personal memory of each particle, encoded through the additional state variable Y_t^i . It can be thought of as an approximation to the in-time best position $\arg \min_{\tau \leq t} \mathcal{E}(X_\tau^i)$ seen by the respective particle. Its dynamics is encoded in Equation (1.2b).

In this section we analyze (1.2) in the large particle limit, i.e., through its mean-field limit (1.8).

3.1 Well-Posedness of PSO with Memory Effects

Ensured by a sufficiently regularized implementation of the local best position \bar{Y} , we can show the well-posedness of the mean-field PSO dynamics (1.8), respectively, the associated Vlasov-Fokker-Planck equation (1.7). As regards uniqueness, it does not seem straightforward to extend the standard proof technique to the present setting due to the way the memory effects are implemented in (1.2b) and (1.8b). Therefore, in what follows, we merely prove existence of solutions and leave the development of a suitably modified proof technique for future research, see also Remark 8.

Theorem 3 *Let \mathcal{E} satisfy Assumptions A1–A3. Moreover, let $m, \gamma, \lambda_1, \lambda_2, \sigma_1, \sigma_2, \alpha, \beta, \theta, \kappa, T > 0$. If $(\bar{X}_0, \bar{Y}_0, \bar{V}_0)$ is distributed according to $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$, then the nonlinear SDE (1.8) admits a strong solution up to time T with $\mathcal{C}([0, T], \mathbb{R}^d) \times \mathcal{C}([0, T], \mathbb{R}^d) \times \mathcal{C}([0, T], \mathbb{R}^d)$ -valued paths. The associated law f has regularity $\mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{3d}))$ and is a weak solution to the Vlasov-Fokker-Planck equation (1.7). In particular,*

$$\sup_{t \in [0, T]} \mathbb{E}[|\bar{X}_t|^4 + |\bar{Y}_t|^4 + |\bar{V}_t|^4] \leq \left(1 + 3\mathbb{E}[|\bar{X}_0|^4 + |\bar{Y}_0|^4 + |\bar{V}_0|^4]\right) e^{CT} \quad (3.1)$$

for some constant $C > 0$ depending only on $m, \gamma, \lambda_1, \lambda_2, \sigma_1, \sigma_2, \alpha, \beta, \theta, \kappa, c_{\mathcal{E}}, R$ and $L_{\mathcal{E}}$.

Proof sketch The proof follows the steps taken in [8, Theorems 3.1, 3.2].

Step 1: For a given function $u \in \mathcal{C}([0, T], \mathbb{R}^d)$ and an initial measure $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$, according to standard SDE theory [2, Chapter 6], we can uniquely solve the auxiliary SDE

$$\begin{aligned} d\tilde{X}_t &= \tilde{V}_t dt, \\ d\tilde{Y}_t &= \kappa(\tilde{X}_t - \tilde{Y}_t) S^{\beta, \theta}(\tilde{X}_t, \tilde{Y}_t) dt, \\ m d\tilde{V}_t &= -\gamma \tilde{V}_t dt + \lambda_1(\tilde{Y}_t - \tilde{X}_t) dt + \lambda_2(u_t - \tilde{X}_t) dt + \sigma_1 D(\tilde{Y}_t - \tilde{X}_t) dB_t^1 \\ &\quad + \sigma_2 D(u_t - \tilde{X}_t) dB_t^2, \end{aligned}$$

with initial condition $(\tilde{X}_0, \tilde{Y}_0, \tilde{V}_0) \sim f_0$ as, due to the smoothness of $S^{\beta, \theta}$ and Assumptions A2 and A3, the coefficients are locally Lipschitz and have at most linear growth. This induces $\tilde{f}_t = \text{Law}(\tilde{X}_t, \tilde{Y}_t, \tilde{V}_t)$. Moreover, the regularity of $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$ allows for a moment estimate of the form (3.1) and thus $\tilde{f} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{3d}))$, see, e.g. [2, Chapter 7]. In what follows, $\tilde{\rho}_Y$ denotes the spatial local best marginal of \tilde{f} , i.e., $\tilde{\rho}_Y(t, \cdot) = \iint_{\mathbb{R}^{2d}} d\tilde{f}(t, x, \cdot, v)$.

Step 2: Let us now define, for some constant $C > 0$, the test function space

$$\begin{aligned} \mathcal{C}_*^2(\mathbb{R}^{3d}) &:= \{\phi \in \mathcal{C}^2(\mathbb{R}^{3d}) : |\nabla_v \phi| \leq C(1 + |x| + |y| + |v|) \\ &\quad \text{and } \sup_{k=1, \dots, d} \left\| \partial_{v_k}^2 \phi \right\|_{\infty} < \infty\}. \end{aligned} \quad (3.2)$$

For some $\phi \in C_*^2(\mathbb{R}^{3d})$, by the Itô-Doeblin formula, we derive

$$\begin{aligned} d\phi &= \nabla_x \phi \cdot \tilde{V}_t dt + \kappa \nabla_y \phi \cdot (\tilde{X}_t - \tilde{Y}_t) S^{\beta, \theta}(\tilde{X}_t, \tilde{Y}_t) dt \\ &\quad + \nabla_v \phi \cdot \left(-\frac{\gamma}{m} \tilde{V}_t + \frac{\lambda_1}{m} (\tilde{Y}_t - \tilde{X}_t) + \frac{\lambda_2}{m} (u_t - \tilde{X}_t) \right) dt \\ &\quad + \frac{1}{2} \sum_{k=1}^d \partial_{v_k v_k}^2 \phi \left(\frac{\sigma_1^2}{m^2} (\tilde{Y}_t - \tilde{X}_t)_k^2 + \frac{\sigma_2^2}{m^2} (u_t - \tilde{X}_t)_k^2 \right) dt \\ &\quad + \nabla_v \phi \cdot \left(\frac{\sigma_1}{m} D(\tilde{Y}_t - \tilde{X}_t) dB_t^1 + \frac{\sigma_2}{m} D(u_t - \tilde{X}_t) dB_t^2 \right), \end{aligned}$$

where we mean $\phi(\tilde{X}_t, \tilde{Y}_t, \tilde{V}_t)$ whenever we write ϕ . After taking the expectation, applying Fubini’s theorem and observing that the stochastic integrals vanish due to the definition of the test function space $C_*^2(\mathbb{R}^{3d})$ and the regularity (3.1), we observe that $\tilde{f} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{3d}))$ satisfies the Vlasov-Fokker-Planck equation

$$\begin{aligned} \frac{d}{dt} \iiint_{\mathbb{R}^{3d}} \phi d\tilde{f}_t &= \iiint_{\mathbb{R}^{3d}} v \cdot \nabla_x \phi d\tilde{f}_t + \iiint_{\mathbb{R}^{3d}} \kappa(x - y) S^{\beta, \theta}(x, y) \cdot \nabla_y \phi d\tilde{f}_t \\ &\quad - \iiint_{\mathbb{R}^{3d}} \left(\frac{\gamma}{m} v + \frac{\lambda_1}{m} (x - y) + \frac{\lambda_2}{m} (x - u_t) \right) \cdot \nabla_v \phi d\tilde{f}_t \\ &\quad + \iiint_{\mathbb{R}^{3d}} \sum_{k=1}^d \left(\frac{\sigma_1^2}{2m^2} (x - y)_k^2 + \frac{\sigma_2^2}{2m^2} (x - u_t)_k^2 \right) \cdot \partial_{v_k v_k}^2 \phi d\tilde{f}_t. \end{aligned} \tag{3.3}$$

Step 3: Setting $\mathcal{T}u := y_\alpha(\tilde{\rho}_Y) \in \mathcal{C}([0, T], \mathbb{R}^d)$ provides the self-mapping property of the map

$$\mathcal{T} : \mathcal{C}([0, T], \mathbb{R}^d) \rightarrow \mathcal{C}([0, T], \mathbb{R}^d), \quad u \mapsto \mathcal{T}u = y_\alpha(\tilde{\rho}_Y),$$

which is compact as a consequence of the stability estimate $|y_\alpha(\tilde{\rho}_{Y,t}) - y_\alpha(\tilde{\rho}_{Y,s})|_2 \lesssim W_2(\tilde{\rho}_{Y,t}, \tilde{\rho}_{Y,s})$ for $\tilde{\rho}_{Y,t}, \tilde{\rho}_{Y,s} \in \mathcal{P}_4(\mathbb{R}^d)$, see, e.g., [8, Lemma 3.2], and the Hölder-1/2 continuity of the Wasserstein-2 distance $W_2(\tilde{\rho}_{Y,t}, \tilde{\rho}_{Y,s})$.

Step 4: Then, for $u = \vartheta \mathcal{T}u$ with $\vartheta \in [0, 1]$, there exists $\tilde{f} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{3d}))$ satisfying (3.3) with marginal $\tilde{\rho}_Y$ such that $u_t = \vartheta y_\alpha(\tilde{\rho}_{Y,t})$. For such u , a uniform bound can be obtained as of Assumption A3. An application of the Leray-Schauder fixed point theorem provides a solution to (1.8). □

3.2 Convergence of PSO with Memory Effects to a Global Minimizer

Analogously to Sect. 2.2 we define a functional $\mathcal{H}(t)$, which is analyzed in this section to eventually prove its exponential decay and thereby consensus formation at some \tilde{x} close to the global minimizer x^* . In addition to the requirements that the variance $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]$ in the position and the second-order moment of the

velocity $\mathbb{E}[|\bar{V}_t|^2]$ of the averaged particle vanish, we also expect that the particle’s position \bar{X}_t aligns with its personal best position \bar{Y}_t over time, meaning that $\mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2]$ decays to zero. This motivates the definition

$$\begin{aligned} \mathcal{H}(t) := & \left(\frac{\gamma}{2m}\right)^2 |\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + \frac{1}{2}|\bar{V}_t|^2 + \frac{1}{2} \left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) |\bar{X}_t - \bar{Y}_t|^2 \\ & + \frac{\gamma}{2m} \langle \bar{X}_t - \mathbb{E}[\bar{X}_t], \bar{V}_t \rangle + \frac{\gamma}{m} \langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle, \end{aligned} \tag{3.4}$$

whose last two terms are required for technical reasons. Again, by the equivalence established in the following Lemma 4, proving the decay of the Lyapunov function $\mathbb{E}[\mathcal{H}(t)]$ directly entails the decay of $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2 + |\bar{X}_t - \bar{Y}_t|^2]$ with the same rate.

Lemma 4 *The functional $\mathcal{H}(t)$ is equivalent to $|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2 + |\bar{X}_t - \bar{Y}_t|^2$ in the sense that*

$$\begin{aligned} & \frac{1}{2} \left(\frac{\gamma}{2m}\right)^2 |\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + \frac{1}{2}|\bar{V}_t|^2 + \frac{3\lambda_1}{2m} |\bar{X}_t - \bar{Y}_t|^2 \leq \mathcal{H}(t) \\ & \leq \frac{5}{2} \left(\left(\frac{\gamma}{2m}\right)^2 + 1 + \frac{3\lambda_1}{m} + \frac{2\gamma^2}{m^2} \right) \left(|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2 + |\bar{V}_t|^2 + |\bar{X}_t - \bar{Y}_t|^2 \right). \end{aligned} \tag{3.5}$$

We now derive an evolution inequality of the Lyapunov function $\mathbb{E}[\mathcal{H}(t)]$.

Lemma 5 *Let \mathcal{E} satisfy Assumptions A1–A3 and let $(\bar{X}_t, \bar{Y}_t, \bar{V}_t)_{t \geq 0}$ be a solution to the nonlinear SDE (1.8). Then $\mathbb{E}[\mathcal{H}(t)]$ with \mathcal{H} as defined in (3.4) satisfies*

$$\begin{aligned} & \frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] \\ & \leq -\frac{\gamma}{2m} \mathbb{E}[|\bar{V}_t|^2] \\ & \quad - \left(\frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2} - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2} \right) \frac{6e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t))]} \right) \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] \\ & \quad - \left(\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} + \kappa\theta \left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2} \right) - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} - \frac{3\sigma_1^2}{2m^2} \right. \\ & \quad \left. - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} \right) - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2} \right) \frac{12e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t))]} \right) \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2]. \end{aligned} \tag{3.6}$$

Proof Let us write $\delta\bar{X}_t := \bar{X}_t - \mathbb{E}[\bar{X}_t]$ for short and note that the integration by parts formula gives

$$\frac{d}{dt} \mathbb{E}[|\delta\bar{X}_t|^2] = 2\mathbb{E}[\langle \delta\bar{X}_t, \bar{V}_t \rangle]. \tag{3.7}$$

Observe that the stochastic integrals have vanishing expectations as a consequence of [39, Theorem 3.2.1(iii), Definition 3.1.4(iii)] combined with the regularity $f \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ obtained in Theorem 3. An application of the Itô-Doebelin formula and Young’s inequality yields

$$\begin{aligned}
 \frac{d}{dt} \mathbb{E}[|\bar{V}_t|^2] &= -\frac{2\gamma}{m} \mathbb{E}[|\bar{V}_t|^2] + \frac{2\lambda_1}{m} \mathbb{E}[\langle \bar{V}_t, \bar{Y}_t - \bar{X}_t \rangle] \\
 &\quad + \frac{2\lambda_2}{m} \mathbb{E}[\langle \bar{V}_t, y_\alpha(\rho_{Y,t}) - \bar{X}_t \rangle] + \frac{\sigma_1^2}{m^2} \mathbb{E}[|\bar{Y}_t - \bar{X}_t|^2] \\
 &\quad + \frac{\sigma_2^2}{m^2} \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{X}_t|^2] \\
 &\leq -\left(\frac{2\gamma}{m} - \frac{\lambda_2}{\varepsilon m}\right) \mathbb{E}[|\bar{V}_t|^2] + \frac{\sigma_1^2}{m^2} \mathbb{E}[|\bar{Y}_t - \bar{X}_t|^2] \\
 &\quad + \left(\frac{\varepsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right) \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{X}_t|^2] \\
 &\quad - \frac{2\lambda_1}{m} \mathbb{E}[\langle \bar{V}_t, \bar{X}_t - \bar{Y}_t \rangle], \quad \forall \varepsilon > 0.
 \end{aligned}
 \tag{3.8}$$

Again by employing the Itô-Doebelin formula we obtain

$$\begin{aligned}
 \frac{d}{dt} \mathbb{E}[\langle \delta \bar{X}_t, \bar{V}_t \rangle] &= \mathbb{E}[|\bar{V}_t|^2] - (\mathbb{E}[\bar{V}_t])^2 - \frac{\gamma}{m} \mathbb{E}[\langle \delta \bar{X}_t, \bar{V}_t \rangle] + \frac{\lambda_1}{m} \mathbb{E}[\langle \delta \bar{X}_t, \bar{Y}_t - \bar{X}_t \rangle] \\
 &\quad + \frac{\lambda_2}{m} \mathbb{E}[\langle \delta \bar{X}_t, y_\alpha(\rho_{Y,t}) - \bar{X}_t \rangle] \\
 &\leq \mathbb{E}[|\bar{V}_t|^2] - \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta \bar{X}_t|^2] \\
 &\quad + \frac{\lambda_1}{m} \mathbb{E}[\langle \delta \bar{X}_t, (\bar{Y}_t - y_\alpha(\rho_{Y,t})) - (\bar{X}_t - \mathbb{E}[\bar{X}_t]) \rangle] \\
 &\quad + \frac{\lambda_2}{m} \mathbb{E}[\langle \delta \bar{X}_t, \mathbb{E}[\bar{X}_t] - \bar{X}_t \rangle] \\
 &= \mathbb{E}[|\bar{V}_t|^2] - \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta \bar{X}_t|^2] - \frac{\lambda_1 + \lambda_2}{m} \mathbb{E}[|\delta \bar{X}_t|^2] \\
 &\quad + \frac{\lambda_1}{m} \mathbb{E}[\langle \delta \bar{X}_t, \bar{Y}_t - y_\alpha(\rho_{Y,t}) \rangle] \\
 &\leq \mathbb{E}[|\bar{V}_t|^2] - \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta \bar{X}_t|^2] - \frac{\lambda_1 + 2\lambda_2}{2m} \mathbb{E}[|\delta \bar{X}_t|^2] \\
 &\quad + \frac{\lambda_1}{2m} \mathbb{E}[|\bar{Y}_t - y_\alpha(\rho_{Y,t})|^2],
 \end{aligned}$$

where, for the second line, we used the identity (3.7) and that $\mathbb{E}[\langle \delta \bar{X}_t, \mathbf{C} \rangle] = 0$, whenever $\mathbf{C} \in \mathbb{R}^d$ is constant, allowing to expand the expression in the way done. We

now rearrange the previous inequality to get

$$\begin{aligned} \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[|\delta \bar{X}_t|^2] + \frac{d}{dt} \mathbb{E}[\langle \delta \bar{X}_t, \bar{V}_t \rangle] &\leq \mathbb{E}[|\bar{V}_t|^2] - \frac{\lambda_1 + 2\lambda_2}{2m} \mathbb{E}[|\delta \bar{X}_t|^2] \\ &\quad + \frac{\lambda_1}{2m} \mathbb{E}[|\bar{Y}_t - y_\alpha(\rho_{Y,t})|^2]. \end{aligned} \tag{3.9}$$

Next, using the Itô-Doeblin formula, we compute

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] &= 2\mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, \bar{V}_t - \kappa(\bar{X}_t - \bar{Y}_t) S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) \rangle] \\ &\leq 2\mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle] - 2\kappa\theta \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2], \end{aligned} \tag{3.10}$$

where the last step follows from the fact that $\theta < S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) < 2 + \theta < 4$. And lastly, the Itô-Doeblin formula and Young’s inequality allow to bound

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle] &= -\frac{\gamma}{m} \mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle] - \frac{\lambda_1 + \lambda_2}{m} \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] + \frac{\lambda_2}{m} \mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, y_\alpha(\rho_{Y,t}) - \bar{Y}_t \rangle] \\ &\quad + \mathbb{E}[\langle \bar{V}_t - \kappa(\bar{X}_t - \bar{Y}_t) S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t), \bar{V}_t \rangle] \\ &\leq -\frac{\gamma}{m} \mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle] - \frac{\lambda_1 + \lambda_2}{m} \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] + \frac{\lambda_2^2}{2m\lambda_1} \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \\ &\quad + \frac{\lambda_1}{2m} \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{Y}_t|^2] \\ &\quad + \mathbb{E}[|\bar{V}_t|^2] + \frac{1}{2} \mathbb{E}[|\bar{V}_t|^2] + 8\kappa^2 \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \\ &= -\left(\frac{\lambda_1 + \lambda_2}{m} - 8\kappa^2 - \frac{\lambda_2^2}{2m\lambda_1}\right) \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] + \frac{3}{2} \mathbb{E}[|\bar{V}_t|^2] + \frac{\lambda_1}{2m} \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{Y}_t|^2] \\ &\quad - \frac{\gamma}{m} \mathbb{E}[\langle \bar{X}_t - \bar{Y}_t, \bar{V}_t \rangle]. \end{aligned} \tag{3.11}$$

We now collect the bounds (3.8), (3.9), (3.10), and (3.11) to show

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] &\leq -\left(\frac{3\gamma}{m} - \frac{3\lambda_2}{2\epsilon m} - \frac{\gamma}{2m} - \frac{3\gamma}{2m}\right) \mathbb{E}[|\bar{V}_t|^2] - \frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2} \mathbb{E}[|\delta \bar{X}_t|^2] \\ &\quad - \left(\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} + \kappa\theta\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) - \frac{3\sigma_1^2}{2m^2}\right) \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \\ &\quad + \frac{3}{2} \left(\frac{\epsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right) \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{X}_t|^2] + \frac{3\lambda_1\gamma}{(2m)^2} \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{Y}_t|^2] \\ &\leq -\left(\frac{\gamma}{m} - \frac{3\lambda_2}{2\epsilon m}\right) \mathbb{E}[|\bar{V}_t|^2] - \frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2} \mathbb{E}[|\delta \bar{X}_t|^2] \\ &\quad - \left(\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} + \kappa\theta\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) - \frac{3\sigma_1^2}{2m^2} - 3\left(\frac{\epsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right)\right) \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \end{aligned}$$

$$+ \left(3 \left(\frac{\varepsilon \lambda_2}{m} + \frac{\sigma_2^2}{m^2} \right) + \frac{3\lambda_1 \gamma}{(2m)^2} \right) \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{Y}_t|^2].$$

Recalling the computation (2.13) yields the bound

$$\begin{aligned} \mathbb{E}[|\bar{Y}_t - y_\alpha(\rho_{Y,t})|^2] &\leq 2e^{-\alpha \underline{\mathcal{E}}} \frac{\mathbb{E}[|\delta \bar{Y}_t|^2]}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_t))]} \\ &\leq 2e^{-\alpha \underline{\mathcal{E}}} \frac{6\mathbb{E}[|\bar{Y}_t - \bar{X}_t|^2] + 3\mathbb{E}[|\delta \bar{X}_t|^2]}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_t))]}, \end{aligned} \tag{3.12}$$

where we inserted $\pm \bar{X}_t$ and $\pm \mathbb{E}[\bar{X}_t]$ in the second step and used that $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ as well as Jensen’s inequality. Combining the last two bounds and choosing $\varepsilon = (3\lambda_2)/\gamma$ we obtain (3.6) as desired. \square

Remark 6 The exponential decay of $\mathbb{E}[\mathcal{H}(t)]$ it obtained by choosing the parameters of PSO in a manner which ensures the negativity of the prefactors of $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]$ and $\mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2]$ in Inequality (3.6). This may be achieved by choosing for any fixed time t , given α and arbitrary $\theta, \sigma_1, \sigma_2, \gamma > 0$,

$$\begin{aligned} \lambda_1 &> \frac{3\sigma_1^2}{2\gamma}, \quad \lambda_2 > 6 \max \left\{ \frac{D_t^Y \lambda_1}{4}, \frac{(1 + D_t^Y)\sigma_2^2}{\gamma} \right\}, \quad \kappa > \frac{3\lambda_2^2(1 + D_t^Y)}{\gamma\theta\lambda_1}, \\ \text{and } m &< \min \left\{ \frac{\gamma\theta}{16\kappa}, \frac{\lambda_1\gamma^2}{18D_t^Y\lambda_2^2} \right\}, \end{aligned}$$

where we abbreviate $D_t^Y = 12e^{-\alpha \underline{\mathcal{E}}}/\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_t))]$.

In our main theorem on convergence of the PSO dynamics with memory mechanisms to the global minimizer x^* we again ensure that the parameter can be chosen once at initialization time.

Theorem 4 *Let \mathcal{E} satisfy Assumptions A1–A4 and let $(\bar{X}_t, \bar{V}_t)_{t \geq 0}$ be a solution to the nonlinear SDE (1.8). Moreover, let us assume the well-preparation of the initial datum \bar{X}_0 and \bar{V}_0 in the sense that*

P1 $\mu_1 > 0$ with

$$\mu_1 := \frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2} - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{4m^2} \right) \frac{12e^{-\alpha \underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_0))]},$$

P2 $\mu_2 > 0$ with

$$\begin{aligned} \mu_2 := &\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} + \kappa\theta \left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2} \right) - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} - \frac{3\sigma_1^2}{2m^2} \\ &- \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} \right) - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2} \right) \frac{24e^{-\alpha \underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_0))]}, \end{aligned}$$

P3 it holds

$$\begin{aligned} & \left(\frac{\alpha \kappa m}{\lambda_1 \chi} \left(C_{\mathcal{E}} + 2\alpha^2 \right) + \frac{24C_{\mathcal{E}}^2 \kappa}{\alpha \chi^3} \right) \frac{\mathbb{E}[\mathcal{H}(0)]}{\mathbb{E}[\exp(-\alpha(\mathcal{E}(\bar{Y}_0) - \underline{\mathcal{E}}))]} \\ & + \frac{6\kappa}{\alpha \chi} \frac{\mathbb{E}[|\nabla \mathcal{E}(\bar{X}_0)|^2]}{\mathbb{E}[\exp(-\alpha(\mathcal{E}(\bar{Y}_0) - \underline{\mathcal{E}}))]} < \frac{3}{32} \end{aligned}$$

where

$$\chi := \frac{2}{5} \frac{\min\{\gamma/(2m), \mu_1, \mu_2\}}{(\gamma/(2m))^2 + 1 + 3\lambda_1/m + 2(\gamma/m)^2}.$$

Then $\mathbb{E}[\mathcal{H}(t)]$ with \mathcal{H} as defined in Equation (3.4) converges exponentially fast with rate χ to 0 as $t \rightarrow \infty$. Moreover, there exists some \tilde{x} , which may depend on α and f_0 , such that $\mathbb{E}[\bar{X}_t] \rightarrow \tilde{x}$ and $y_{\alpha}(\rho_{Y,t}) \rightarrow \tilde{x}$ exponentially fast with rate $\chi/2$ as $t \rightarrow \infty$. Eventually, for any given accuracy $\varepsilon > 0$, there exists $\alpha_0 > 0$, which may depend on the dimension d , such that for all $\alpha > \alpha_0$, \tilde{x} satisfies

$$\mathcal{E}(\tilde{x}) - \underline{\mathcal{E}} \leq \varepsilon.$$

If \mathcal{E} additionally satisfies Assumption A5, we additionally have $|\tilde{x} - x^*| \leq \varepsilon^{\nu}/\eta$.

Remark 7 By replacing D_t^Y with $2D_0^Y$ in the parameter choices of Remark 6, the well-preparation of the parameters as in Conditions P1 and P2 can be ensured.

In analogy to Remark 4, Condition P3 guarantees the well-preparation of the initialization.

Proof of Theorem 4 Let us define the time horizon

$$T := \inf \left\{ t \geq 0 : \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_t))] < \frac{1}{2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_0))] \right\} \quad \text{with } \inf \emptyset = \infty.$$

Obviously, by continuity, $T > 0$. We claim that $T = \infty$, which we prove by contradiction in the following. Therefore, assume $T < \infty$. Then, for $t \in [0, T]$, noting that $\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_t))] \geq \mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_0))]/2$, we observe that the prefactors of $\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2]$ and $\mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2]$ in Lemma 5 are upper bounded by $-\mu_1$ and $-\mu_2$, respectively. Lemma 5 then provides an upper bound for the time derivative of the functional $\mathbb{E}[\mathcal{H}(t)]$,

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] & \leq -\frac{\gamma}{2m} \mathbb{E}[|\bar{V}_t|^2] - \mu_1 \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] - \mu_2 \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \\ & \leq -\min \left\{ \frac{\gamma}{2m}, \mu_1, \mu_2 \right\} \left(\mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] + \mathbb{E}[|\bar{V}_t|^2] + \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \right) \quad (3.13) \\ & \leq -\frac{2}{5} \frac{\min\{\gamma/(2m), \mu_1, \mu_2\}}{(\gamma/(2m))^2 + 1 + 3\lambda_1/m + 2\gamma^2/m^2} \mathbb{E}[\mathcal{H}(t)] =: -\chi \mathbb{E}[\mathcal{H}(t)], \end{aligned}$$

where we made use of the upper bound of (3.5) as in Lemma 4 in the last inequality. The rate χ is defined implicitly and it is straightforward to check that $0 < \chi < \gamma/m$,

where the positivity of χ follows from the well-preparation conditions **P1** and **P2** of the initialization. Grönwall’s inequality implies

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t). \tag{3.14}$$

We now investigate the evolution of the functional $\mathcal{Y}(t) := \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t))]$. The Itô-Doebelin formula yields

$$\begin{aligned} \frac{d}{dt}\mathcal{Y}(t) &= -\alpha\kappa\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t)) \langle \nabla\mathcal{E}(\bar{Y}_t), (\bar{X}_t - \bar{Y}_t)S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) \rangle] \\ &= -\alpha\kappa\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t)) \langle \nabla\mathcal{E}(\bar{Y}_t) - \nabla\mathcal{E}(\bar{X}_t), (\bar{X}_t - \bar{Y}_t)S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) \rangle] \\ &\quad - \alpha\kappa\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t)) \langle \nabla\mathcal{E}(\bar{X}_t), (\bar{X}_t - \bar{Y}_t)S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) \rangle] \\ &\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}C_{\mathcal{E}}\mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] - 4\alpha\kappa e^{-\alpha\mathcal{E}}\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_t)||\bar{X}_t - \bar{Y}_t|], \end{aligned} \tag{3.15}$$

where the last step follows from Cauchy-Schwarz inequality and uses Assumption **A4** and $S^{\beta,\theta}(\bar{X}_t, \bar{Y}_t) < 4$. Now firstly notice that $\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_t)||\bar{X}_t - \bar{Y}_t|] \leq e^{(\chi/2)t}\alpha^2\mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] + e^{-(\chi/2)t}/\alpha^2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_t)|^2]$ by Young’s inequality. Secondly, using again Assumption **A4** in the first inequality, we have

$$\begin{aligned} \mathbb{E}[|\nabla\mathcal{E}(\bar{X}_t)|^2] &= \mathbb{E}\left[\left|\nabla\mathcal{E}(\bar{X}_0) + \int_0^t \nabla^2\mathcal{E}(\bar{X}_s)\bar{V}_s ds\right|^2\right] \\ &\leq 2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] + 2C_{\mathcal{E}}^2t \int_0^t \mathbb{E}[|\bar{V}_s|^2] ds \\ &\leq 2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] + 4C_{\mathcal{E}}^2t \int_0^t \mathbb{E}[\mathcal{H}(s)] ds \\ &\leq 2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] + 4C_{\mathcal{E}}^2t\mathbb{E}[\mathcal{H}(0)] \int_0^t \exp(-\chi s) ds \\ &= 2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] + 4C_{\mathcal{E}}^2t\mathbb{E}[\mathcal{H}(0)]\frac{1}{\chi}(1 - \exp(-\chi t)), \end{aligned}$$

where the next-to-last step uses the explicit bound in (3.14). Using the two latter observations together with the fact that $\mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] \leq 2m/(3\lambda_1)\mathbb{E}[\mathcal{H}(t)]$ we can continue (3.15) as follows

$$\begin{aligned} \frac{d}{dt}\mathcal{Y}(t) &\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}\left(C_{\mathcal{E}} + \exp\left(\frac{\chi}{2}t\right)\alpha^2\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(t)] \\ &\quad - \frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\exp\left(-\frac{\chi}{2}t\right)\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_t)|^2] \\ &\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}\left(C_{\mathcal{E}} + \exp\left(\frac{\chi}{2}t\right)\alpha^2\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(0)]\exp(-\chi t) \\ &\quad - \frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\exp\left(-\frac{\chi}{2}t\right)\left(2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] + 4C_{\mathcal{E}}^2t\mathbb{E}[\mathcal{H}(0)]\frac{1}{\chi}(1 - \exp(-\chi t))\right) \end{aligned}$$

$$\begin{aligned} &\geq -4\alpha\kappa e^{-\alpha\varepsilon} \left(C_\varepsilon \exp(-\chi t) + \exp\left(-\frac{\chi}{2}t\right) \alpha^2 \right) \frac{2m}{3\lambda_1} \mathbb{E}[\mathcal{H}(0)] \\ &\quad - \frac{4}{\alpha} \kappa e^{-\alpha\varepsilon} \exp\left(-\frac{\chi}{2}t\right) \left(2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] + \frac{4C_\varepsilon^2 t}{\chi} \mathbb{E}[\mathcal{H}(0)] \right). \end{aligned} \tag{3.16}$$

By integrating (3.16) we obtain for all $t \in [0, T]$

$$\begin{aligned} \mathcal{Y}(t) &\geq \mathcal{Y}(0) - 4\alpha\kappa e^{-\alpha\varepsilon} \left(\frac{C_\varepsilon}{\chi} + \frac{2\alpha^2}{\chi} \right) \frac{2m}{3\lambda_1} \mathbb{E}[\mathcal{H}(0)] \\ &\quad - \frac{4}{\alpha} \kappa e^{-\alpha\varepsilon} \left(2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] \frac{2}{\chi} + \frac{16C_\varepsilon^2}{\chi^3} \mathbb{E}[\mathcal{H}(0)] \right). \end{aligned}$$

Recalling the definition of \mathcal{Y} and employing Condition P3, we can deduce that for all $t \in [0, T]$ it holds

$$\begin{aligned} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t))] &\geq \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_0))] - 4\alpha\kappa e^{-\alpha\varepsilon} \left(\frac{C_\varepsilon}{\chi} + \frac{2\alpha^2}{\chi} \right) \frac{2m}{3\lambda_1} \mathbb{E}[\mathcal{H}(0)] \\ &\quad - \frac{4}{\alpha} \kappa e^{-\alpha\varepsilon} \left(2\mathbb{E}[|\nabla\mathcal{E}(\bar{X}_0)|^2] \frac{2}{\chi} + \frac{16C_\varepsilon^2}{\chi^3} \mathbb{E}[\mathcal{H}(0)] \right) \\ &> \frac{3}{4} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_0))], \end{aligned}$$

which entails that there exists $\delta > 0$ such that $\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t))] \geq \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_0))]/2$ in $[T, T + \delta]$ as well, contradicting the definition of T and therefore showing the claim $T = \infty$.

As a consequence of (3.14) we have

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t) \quad \text{and} \quad \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_t))] \geq \frac{1}{2} \mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{Y}_0))] \tag{3.17}$$

for all $t \geq 0$. In particular, by means of Lemma 4, for a suitable generic constant $C > 0$, we infer

$$\begin{aligned} \mathbb{E}[|\bar{X}_t - \mathbb{E}[\bar{X}_t]|^2] &\leq C \exp(-\chi t), \quad \mathbb{E}[|\bar{V}_t|^2] \leq C \exp(-\chi t), \\ \text{and} \quad \mathbb{E}[|\bar{X}_t - \bar{Y}_t|^2] &\leq C \exp(-\chi t). \end{aligned} \tag{3.18}$$

Moreover, with Jensen's inequality,

$$\left| \frac{d}{dt} \mathbb{E}[\bar{X}_t] \right| \leq \mathbb{E}[|\bar{V}_t|] \leq C \exp(-\chi t/2) \rightarrow 0 \quad \text{as } t \rightarrow \infty,$$

showing that $\mathbb{E}[\bar{X}_t] \rightarrow \tilde{x}$ for some $\tilde{x} \in \mathbb{R}^d$, which may depend on α and f_0 . According to (3.18), $\bar{X}_t \rightarrow \tilde{x}$ as well as $\bar{Y}_t \rightarrow \tilde{x}$ in mean-square. Moreover, by reusing the inequality (3.12) we get

$$\mathbb{E}[|\bar{Y}_t - y_\alpha(\rho_{Y,t})|^2] \leq 4e^{-\alpha \mathcal{E}} \frac{6\mathbb{E}[|\bar{Y}_t - \bar{X}_t|^2] + 3\mathbb{E}[|\bar{X}_t - \mathbb{E}\bar{X}_t|^2]}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\bar{Y}_0))]} \leq C \exp(-\chi t)$$

showing $y_\alpha(\rho_{Y,t}) \rightarrow \tilde{x}$, since

$$\begin{aligned} |y_\alpha(\rho_{Y,t}) - \tilde{x}|^2 &\leq 4\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \bar{Y}_t|^2] + 4\mathbb{E}[|\bar{Y}_t - \bar{X}_t|^2] \\ &\quad + 4\mathbb{E}[|\bar{X}_t - \mathbb{E}\bar{X}_t|^2] + 4|\mathbb{E}\bar{X}_t - \tilde{x}|^2 \rightarrow 0 \quad \text{as } t \rightarrow \infty. \end{aligned}$$

The remainder of the proof follows the lines of the proof of Theorem 2, replacing merely \bar{X}_t with \bar{Y}_t . \square

4 A Holistic Convergence Statement of PSO Without Memory Effects

In Sects. 2 and 3 we analyzed the macroscopic behavior of PSO without and with memory effects in the mean-field regime. For this purpose we introduced the with (1.2) and (2.1) associated self-consistent mono-particle processes (1.8) and (2.3), for which we then established convergence guarantees under the in Theorems 2 and 4 specified assumptions. However, in order to be able to infer therefrom the optimization capabilities of the numerically implemented PSO method, a quantitative estimate on the approximation quality of the interacting particle system by the corresponding mean-field dynamics is necessary.

4.1 On the Mean-Field Approximation of PSO Without Memory Effects

The following theorem provides a probabilistic quantitative estimate on the mean-field approximation for PSO without memory effects. Notably, the result does not suffer from the curse of dimensionality.

Theorem 5 *Let $T > 0$, $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$ and let $N \in \mathbb{N}$ be fixed. Moreover, let \mathcal{E} obey Assumptions A1–A4. We denote by $((X_t^i, V_t^i)_{t \geq 0})_{i=1, \dots, N}$ the solution to system (2.1) and let $((\bar{X}_t^i, \bar{V}_t^i)_{t \geq 0})_{i=1, \dots, N}$ be N independent copies of the solution to the mean-field*

dynamics (2.3). Then it holds

$$\begin{aligned} \mathbb{P}(\Omega_M) &= \mathbb{P}\left(\sup_{t \in [0, T]} \left[\frac{1}{N} \sum_{i=1}^N \max\{|X_t^i|^4 + |V_t^i|^4, |\bar{X}_t^i|^4 + |\bar{V}_t^i|^4\} \right] \leq M\right) \\ &\geq 1 - \frac{2K}{M}, \end{aligned} \tag{4.1}$$

where $K = K(\gamma/m, \lambda/m, \sigma/m, T, \mathcal{E})$ is a constant, which is in particular independent of N and d .

Furthermore, if the processes share the initial data as well as the Brownian motion paths $(B_t^i)_{t \geq 0}$ for all $i = 1, \dots, N$, then we have a probabilistic mean-field approximation of the form

$$\max_{i=1, \dots, N} \sup_{t \in [0, T]} \mathbb{E} \left[|X_t^i - \bar{X}_t^i|^2 + |V_t^i - \bar{V}_t^i|^2 \mid \Omega_M \right] \leq C_{\text{MFA}} N^{-1} \tag{4.2}$$

with a constant $C_{\text{MFA}} = C_{\text{MFA}}(\alpha, \gamma/m, \lambda/m, \sigma/m, T, \mathcal{E}, K, M)$, which is in particular independent of N and d .

Proof The proof is based on the arguments of [18, Section 3.3] about the mean-field approximation of CBO. First we compute a bound for $\mathbb{E}[\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N \max\{|X_t^i|^4 + |V_t^i|^4, |\bar{X}_t^i|^4 + |\bar{V}_t^i|^4\}]$, which is then used to derive a mean-field approximation for PSO conditioned on the set Ω_M of uniformly bounded processes.

Step 1: Using standard inequalities and Jensen’s inequality allows to derive the bound

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in [0, T]} |X_t^i|^4 \right] &\lesssim \mathbb{E}[|X_0^i|^4] + \mathbb{E} \left[\sup_{t \in [0, T]} \left| \int_0^t V_s^i ds \right|^4 \right] \\ &\leq C \left(\mathbb{E}[|X_0^i|^4] + \mathbb{E} \left[\int_0^T |V_s^i|^4 ds \right] \right) \end{aligned} \tag{4.3}$$

with $C = C(T)$. For the velocities V_t^i we first note that

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in [0, T]} |V_t^i|^4 \right] &\lesssim \mathbb{E}[|V_0^i|^4] + \left(\frac{\gamma}{m}\right)^4 \mathbb{E} \left[\sup_{t \in [0, T]} \left| \int_0^t V_s^i ds \right|^4 \right] \\ &\quad + \left(\frac{\lambda}{m}\right)^4 \mathbb{E} \left[\sup_{t \in [0, T]} \left| \int_0^t (x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i) ds \right|^4 \right] \\ &\quad + \left(\frac{\sigma}{m}\right)^4 \mathbb{E} \left[\sup_{t \in [0, T]} \left| \int_0^t D(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i) dB_s^i \right|^4 \right]. \end{aligned} \tag{4.4}$$

While the two middle terms on the right-hand side of (4.4) can be controlled as before by applying Jensen’s inequality, the last term is treated as follows. Since

$\int_0^t D(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i)dB_s^i$ is a martingale we can apply the Burkholder-Davis-Gundy inequality [47, Chapter IV, Theorem 4.1], which gives

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in [0, T]} \left| \int_0^t D(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i)dB_s^i \right|^4 \right] &\lesssim \sup_{t \in [0, T]} \mathbb{E} \left[\left(\int_0^t |x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i|^2 ds \right)^2 \right] \\ &\leq C \mathbb{E} \left[\int_0^T |x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i|^4 ds \right], \end{aligned} \tag{4.5}$$

where the latter step is again due to Jensen’s inequality and with a constant $C = C(T)$. Utilizing these bounds allows to continue the inequality in (4.4) and to obtain

$$\mathbb{E} \left[\sup_{t \in [0, T]} |V_t^i|^4 \right] \leq C \left(\mathbb{E}[|V_0^i|^4] + \mathbb{E} \left[\int_0^T |X_s^i|^4 + |x_\alpha(\widehat{\rho}_{X,s}^N)|^4 + |V_s^i|^4 ds \right] \right) \tag{4.6}$$

with $C = C(\gamma/m, \lambda/m, \sigma/m, T)$. Since according to [8, Lemma 3.3] it holds

$$\begin{aligned} |x_\alpha(\widehat{\rho}_{X,s}^N)|^2 &\leq \int |x|^2 \frac{\omega_\alpha^\mathcal{E}(x)}{\|\omega_\alpha^\mathcal{E}\|_{L_1(\widehat{\rho}_{X,s}^N)}} d\widehat{\rho}_{X,s}^N(x) \leq b_1 + b_2 \int |x|^2 d\widehat{\rho}_{X,s}^N(y) \\ &= b_1 + b_2 \frac{1}{N} \sum_{i=1}^N |X_s^i|^2 \end{aligned}$$

with $b_1 = 0$ and $b_2 = e^{\alpha(\bar{\mathcal{E}}-\mathcal{E})}$ in the case that \mathcal{E} is bounded, and

$$b_1 = R^2 + b_2^2 \text{ and } b_2 = \frac{2L\mathcal{E} \max\{1, |x^*|^2\}}{c\mathcal{E}} \left(1 + \frac{1}{\alpha c\mathcal{E}R^2} \right)$$

in the case that \mathcal{E} satisfies the coercivity assumption A3, we eventually obtain the upper bound

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in [0, T]} |V_t^i|^4 \right] &\leq C \left(1 + \mathbb{E}[|V_0^i|^4] + \mathbb{E} \left[\int_0^T |X_s^i|^4 + \frac{1}{N} \sum_{j=1}^N |X_s^j|^4 + |V_s^i|^4 ds \right] \right) \end{aligned} \tag{4.7}$$

with $C = C(\gamma/m, \lambda/m, \sigma/m, T, b_1, b_2)$. Adding up (4.3) and (4.7) yields

$$\mathbb{E} \left[\sup_{t \in [0, T]} |X_t^i|^4 + |V_t^i|^4 \right] \leq C \left(1 + \mathbb{E}[|X_0^i|^4 + |V_0^i|^4] + \mathbb{E} \left[\int_0^T |X_s^i|^4 + \frac{1}{N} \sum_{j=1}^N |X_s^j|^2 + |V_s^i|^4 ds \right] \right), \tag{4.8}$$

which, averaged over i , allows to derive the bound

$$\mathbb{E} \left[\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N (|X_t^i|^4 + |V_t^i|^4) \right] \leq C \left(1 + \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N (|X_0^i|^4 + |V_0^i|^4) \right] + \int_0^T \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N (|X_s^i|^4 + |V_s^i|^4) \right] ds \right). \tag{4.9}$$

An application of Grönwall’s inequality ensures that $\mathbb{E} \sup_{t \in [0, T]} \left[\frac{1}{N} \sum_{i=1}^N (|X_t^i|^4 + |V_t^i|^4) \right]$ is bounded independently of N by some constant $K = K(\gamma/m, \lambda/m, \sigma/m, T, b_1, b_2)$. Note, that the constant K does in particular not depend on N or d . With identical arguments for the processes $(\bar{X}_t^i, \bar{V}_t^i)$ an analogous bound can be obtained for $\mathbb{E} \left[\sup_{t \in [0, T]} \frac{1}{N} \sum_{i=1}^N (|\bar{X}_t^i|^4 + |\bar{V}_t^i|^4) \right]$. The first claim of the statement now follows from Markov’s inequality.

Step 2: We define the cutoff function

$$I_M(t) = \begin{cases} 1, & \text{if } \frac{1}{N} \sum_{i=1}^N \max \{ |X_s^i|^4 + |V_s^i|^4, |\bar{X}_s^i|^4 + |\bar{V}_s^i|^4 \} \leq M \text{ for all } s \in [0, t], \\ 0, & \text{else,} \end{cases} \tag{4.10}$$

which is a random variable adapted to the natural filtration and satisfying $\mathbb{1}_{\Omega_M} \leq I_M(t)$ pointwise for all $t \in [0, T]$ as well as $I_M(t) = I_M(t)I_M(s)$ for all $s \in [0, t]$. Firstly, for the positions, by using standard inequalities and Jensen’s inequality, we obtain the bound

$$\begin{aligned} \mathbb{E}[|X_t^i - \bar{X}_t^i|^2 I_M(t)] &\lesssim \mathbb{E}[|X_0^i - \bar{X}_0^i|^2] + \mathbb{E} \left[\left| \int_0^t (V_s^i - \bar{V}_s^i) I_M(s) ds \right|^2 \right] \\ &\leq C \left(\mathbb{E}[|X_0^i - \bar{X}_0^i|^2] + \int_0^t \mathbb{E} [|V_s^i - \bar{V}_s^i|^2 I_M(s)] ds \right) \end{aligned} \tag{4.11}$$

with $C = C(T)$. Secondly, for the velocities we have

$$\begin{aligned}
 & \mathbb{E}[|V_t^i - \bar{V}_t^i|^2 I_M(t)] \\
 & \lesssim \mathbb{E}[|V_0^i - \bar{V}_0^i|^2] + \left(\frac{\gamma}{m}\right)^2 \mathbb{E}\left[\left|\int_0^t (V_s^i - \bar{V}_s^i) I_M(s) ds\right|^2\right] \\
 & \quad + \left(\frac{\lambda}{m}\right)^2 \mathbb{E}\left[\left|\int_0^t \left((x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i) - (x_\alpha(\rho_{X,s}) - \bar{X}_s^i)\right) I_M(s) ds\right|^2\right] \\
 & \quad + \left(\frac{\sigma}{m}\right)^2 \mathbb{E}\left[\left|\int_0^t \left(|x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i| - |x_\alpha(\rho_{X,s}) - \bar{X}_s^i|\right) I_M(s) dB_s^i\right|^2\right] \\
 & \leq C \left(\mathbb{E}[|V_0^i - \bar{V}_0^i|^2] + \int_0^t \mathbb{E}[|V_s^i - \bar{V}_s^i|^2 I_M(s)] ds \right. \\
 & \quad \left. + \int_0^t \mathbb{E}\left[\left(|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 + |X_s^i - \bar{X}_s^i|^2\right) I_M(s) \right] ds \right) \tag{4.12}
 \end{aligned}$$

with $C = C(\gamma/m, \lambda/m, \sigma/m, T)$. In the first step of (4.12) we used that the processes (X_t^i, V_t^i) and $(\bar{X}_t^i, \bar{V}_t^i)$ share the Brownian motion paths, and in the second both Itô isometry and Jensen’s inequality. In order to conclude, it remains to control the term $\mathbb{E}\left[|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 I_M(s)\right]$. To do so, in analogy to the definition of $\widehat{\rho}_{X,s}^N$, let us denote by $\bar{\rho}_{X,s}^N$ the empirical measure associated with the processes \bar{X}_s^i , i.e., $\bar{\rho}_{X,s}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\bar{X}_s^i}$. Then, by following the proofs of [8, Lemma 3.2] and [16, Lemma 3.1], and exploiting the boundedness ensured by the multiplication with the random variable $I_M(s)$, we obtain

$$\begin{aligned}
 & \mathbb{E}\left[|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 I_M(s)\right] \\
 & \lesssim \mathbb{E}\left[|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\bar{\rho}_{X,s}^N)|^2 I_M(s)\right] + \mathbb{E}\left[|x_\alpha(\bar{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 I_M(s)\right] \\
 & \leq C \left(\frac{1}{N} \sum_{i=1}^N \mathbb{E}[|X_s^i - \bar{X}_s^i|^2 I_M(s)] + N^{-1} \right) \\
 & \leq C \left(\max_{i=1, \dots, N} \mathbb{E}[|X_s^i - \bar{X}_s^i|^2 I_M(s)] + N^{-1} \right)
 \end{aligned}$$

with $C = C(\alpha, L_{\mathcal{E}}, c_{\mathcal{E}}, |x^*|, M, b_1, b_2)$. Inserting the latter into (4.12), and adding up (4.11) and (4.12) yields

$$\begin{aligned}
 & \mathbb{E}\left[(|X_t^i - \bar{X}_t^i|^2 + |V_t^i - \bar{V}_t^i|^2) I_M(t) \right] \\
 & \leq C \int_0^t \mathbb{E}\left[(|X_s^i - \bar{X}_s^i|^2 + |V_s^i - \bar{V}_s^i|^2) I_M(s) \right] ds \\
 & \quad + \max_{j=1, \dots, N} \mathbb{E}\left[|X_s^j - \bar{X}_s^j|^2 I_M(s) \right] + N^{-1} ds \tag{4.13}
 \end{aligned}$$

with $C = C(\alpha, \gamma/m, \lambda/m, \sigma/m, T, L_{\mathcal{E}}, c_{\mathcal{E}}, |x^*|, M, b_1, b_2)$ and where we used that the processes (X_t^i, V_t^i) and $(\bar{X}_t^i, \bar{V}_t^i)$ share the initial conditions. Lastly, by taking the maximum over i on both sides we get

$$\begin{aligned} & \max_{i=1, \dots, N} \mathbb{E}[(|X_t^i - \bar{X}_t^i|^2 + |V_t^i - \bar{V}_t^i|^2) I_M(t)] \\ & \leq C \int_0^t \mathbb{E} \left[\max_{j=1, \dots, N} \mathbb{E} \left[(|X_s^j - \bar{X}_s^j|^2 + |V_s^j - \bar{V}_s^j|^2) I_M(s) \right] + N^{-1} \right] ds \end{aligned} \tag{4.14}$$

with the C from before. After recalling the definition of the conditional expectation, an application of Grönwall’s inequality concludes the proof. \square

Remark 8 While the first part of Theorem 5 about the uniform in time boundedness of the empirical measures holds mutatis mutandis for the PSO dynamics with memory effects (1.2) and (1.8), it does not seem straightforward to obtain the second part in this setting due to the way the memory effects are implemented in (1.2b) and (1.8b). As a matter of fact, this is due to exactly the same technical reasons why we lack a uniqueness statement in Sect. 3.1. We therefore leave the investigation of this extension to future research, in particular in regard to the question whether a suitably modified proof technique or another implementations of memory effects resolve this issue.

4.2 Convergence of PSO Without Memory Effects in Probability

Combining Theorem 5 with the convergence analysis of the mean-field dynamics (2.1) as described in Theorem 2, as well as a classical result about the numerical approximation of SDEs allows to obtain convergence guarantees with provable polynomial complexity for the numerical PSO method as stated in Theorem 6 below. Let us, for the reader’s convenience, recall from [21, Section 6] that a possible discretized version of the interacting particle system (2.1) is given by

$$X_{(k+1)\Delta t}^i = X_{k\Delta t}^i + \Delta t V_{(k+1)\Delta t}^i, \tag{4.15a}$$

$$\begin{aligned} V_{(k+1)\Delta t}^i = & \left(\frac{m}{m + \Delta t \gamma} \right) V_{k\Delta t}^i + \left(\frac{\Delta t \lambda}{m + \Delta t \gamma} \right) \left(x_{\alpha}(\widehat{\rho}_{X, k\Delta t}^N) - X_{k\Delta t}^i \right) \\ & + \left(\frac{\sqrt{\Delta t} \sigma}{m + \Delta t \gamma} \right) D \left(x_{\alpha}(\widehat{\rho}_{X, k\Delta t}^N) - X_{k\Delta t}^i \right) B_{k\Delta t}^i \end{aligned} \tag{4.15b}$$

for $k = 0, \dots, K$ and where $((B_{k\Delta t}^i)_{k=1, \dots, K-1})_{i=1, \dots, N}$ are independent, identically distributed standard Gaussian random vectors in \mathbb{R}^d .

Theorem 6 *Let $\epsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Then, under the assumptions of Theorems 2 and 5, it holds for the discretized PSO dynamics (4.15) that*

$$\left| \frac{1}{N} \sum_{i=1}^N X_{K\Delta t}^i - x^* \right|^2 \leq \epsilon_{\text{total}} \tag{4.16}$$

with probability larger than $1 - (\delta + \epsilon_{\text{total}}^{-1}(C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \tilde{\epsilon} + \epsilon^{2\nu}/\eta^2))$. Here, m denotes the order of accuracy of the used discretization scheme. Moreover, besides problem dependent factors and the parameters of the method, the dependence of the constants is as follows. C_{NA} depends linearly on d and N , and exponentially on T . C_{MFA} depends on exponentially on α , T and δ^{-1} . C_{LLN} depends on the moment bound from Theorem 1. Lastly, $\tilde{\epsilon}$ and ϵ are chosen according to Theorem 2.

Remark 9 It is worth emphasizing at this point that the time horizon T in Theorem 6 scales as $\mathcal{O}(\log(\tilde{\epsilon}^{-1})/\chi)$ and therefore logarithmically in the desired accuracy $\tilde{\epsilon}$ as a result of Theorem 2, see also the proof below. This ensures that the constants C_{NA} and C_{MFA} appearing implicitly in the bound (4.16) do not lead to an unfeasible numerical method by requiring extremely small time step sizes Δt and an exceedingly large amount of particles N .

Proof of Theorem 6 The overall error can be decomposed as

$$\begin{aligned} & \mathbb{E} \left[\left| \frac{1}{N} \sum_{i=1}^N X_{K\Delta t}^i - x^* \right|^2 \middle| \Omega_M \right] \\ & \lesssim \mathbb{E} \left[\left| \frac{1}{N} \sum_{i=1}^N (X_{K\Delta t}^i - X_T^i) \right|^2 \right] + \mathbb{E} \left[\left| \frac{1}{N} \sum_{i=1}^N (X_T^i - \bar{X}_T) \right|^2 \middle| \Omega_M \right] \quad (4.17) \\ & \quad + \mathbb{E} \left[\left| \frac{1}{N} \sum_{i=1}^N \bar{X}_T^i - \mathbb{E}[\bar{X}_T] \right|^2 \right] + |\mathbb{E}[\bar{X}_T] - \tilde{x}|^2 + |\tilde{x} - x^*|^2, \end{aligned}$$

where we used that $\mathbb{P}(\Omega_M) \geq (1 - \delta) \geq 1/2$. By means of a classical result about the convergence of numerical schemes for SDEs [43], the first term in (4.17) can be bounded by $C_{\text{NA}}(\Delta t)^m$. For the second term, Theorem 5 gives the estimate $C_{\text{MFA}}N^{-1}$. The third term can be bounded by $C_{\text{LLN}}N^{-1}$ as a consequence of the law of large numbers. Eventually, Theorem 2 allows us to choose $T = \mathcal{O}(\log(\tilde{\epsilon}^{-1})/\chi)$ sufficiently large to reach any prescribed accuracy $\tilde{\epsilon}$ for the next-to-last term as well as $\epsilon^{2\nu}/\eta^2$ for the last term by a suitable choice of α . With these individual bounds we obtain

$$\begin{aligned} & \mathbb{E} \left[\left| \frac{1}{N} \sum_{i=1}^N X_{K\Delta t}^i - x^* \right|^2 \middle| \Omega_M \right] \quad (4.18) \\ & \leq C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \tilde{\epsilon} + \epsilon^{2\nu}/\eta^2. \end{aligned}$$

It now remains to estimate the probability of the set $K_{\epsilon_{\text{total}}}^N \subset \Omega$, where Inequality (4.16) does not hold. By utilizing the conditional version of Markov’s inequality together with

the formerly established bound (4.18), we have

$$\begin{aligned}
 \mathbb{P}(K_{\epsilon_{\text{total}}}^N) &= \mathbb{P}(K_{\epsilon_{\text{total}}}^N \cap \Omega_M) + \mathbb{P}(K_{\epsilon_{\text{total}}}^N \cap \Omega_M^c) \\
 &\leq \mathbb{P}(K_{\epsilon_{\text{total}}}^N \mid \Omega_M) \mathbb{P}(\Omega_M) + \mathbb{P}(\Omega_M^c) \\
 &\leq \frac{C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \tilde{\epsilon} + \epsilon^{2\nu}/\eta^2}{\epsilon_{\text{total}}} + \delta
 \end{aligned}
 \tag{4.19}$$

for a sufficiently large choice of M in (4.1). □

A result in this spirit was first presented for CBO in [18, Theorem 14] and is hereby extended to PSO.

5 Implementation of PSO and Numerical Results

The purpose of this section is twofold. At first, an efficient implementation of PSO is provided, which is particularly suited for high-dimensional optimization problems arising in machine learning. Its performance is then evaluated on a standard benchmark problem, where we investigate the influence of the parameters, and the training of a neural network classifier for handwritten digits. Furthermore, several relevant implementational aspects are discussed, including the computational complexity and scalability, modifications inspired from simulated annealing and evolutionary algorithms, and the numerical stability of the method.

5.1 An Efficient Implementation of PSO

Let us stress that PSO is an extremely versatile, flexible and customizable optimization method, which can be regarded as a black-box optimizer. As a zero-order method it is not reliant on the gradient information and can be even applied to discontinuous objectives, making it inevitably superior to first-order optimization methods in cases where derivatives are computationally infeasible. However, also in machine learning applications, where gradient-based optimizers are considered the state of the art, PSO may be of particular interest in view of vanishing or exploding gradient phenomena.

Typical objective functions appearing in machine learning are of the form

$$\mathcal{E}(x) = \frac{1}{M} \sum_{j=1}^M \mathcal{E}_j(x),
 \tag{5.1}$$

where \mathcal{E}_j is usually the loss of the j th training sample. In order to run the scheme (1.2), frequent evaluations of \mathcal{E} are necessary, which may be computationally intense or even prohibitive in some applications.

Computational complexity: Inspired by mini-batch gradient descent, the authors of [28] developed a random batch method for interacting particle systems, which was employed for CBO in [9]. In the same spirit, we present with Algorithm 1 a

Algorithm 1 Particle swarm optimization (PSO)

Input: Objective \mathcal{E} as in (5.1), time horizon T or number of epochs $\#epochs$, discrete time step size Δt , batch sizes n_N and $n_{\mathcal{E}}$, parameters $m, \gamma, \lambda_1, \lambda_2, \sigma_1, \sigma_2, \alpha, \beta, \theta$ and κ , number of particles N , initialization f_0

Output: Approximation $y_{\alpha}(\widehat{\rho}_{Y,T}^N)$ of the global minimizer x^* of \mathcal{E}

- 1: Generate the particles' initial positions and velocities $(X_0^i, V_0^i)_{i=1,\dots,N}$ according to a common initial law f_0 . Initialize the local best positions $Y_0^i = X_0^i$.
- 2: Ensure that $n_{\mathcal{E}}$ divides M and n_N divides N .
- 3: Convert T into $\#epochs$ or vice versa via $T = \#epochs (M/n_{\mathcal{E}})(N/n_N)\Delta t$. Set $k = 0$ and $epoch = 1$.
- 4: **while** $epoch \leq \#epochs$ and stopping criterion not fulfilled
- 5: Partition $\{1, \dots, M\}$ into batches $\mathcal{B}_k^1, \dots, \mathcal{B}_k^{M/n_{\mathcal{E}}}$ of batch size $n_{\mathcal{E}}$.
- 6: **for** $b = 1, \dots, M/n_{\mathcal{E}}$
- 7: Define the objective function on this batch as

$$\mathcal{E}_{batch}(x) = \frac{1}{n_{\mathcal{E}}} \sum_{j \in \mathcal{B}_k^b} \mathcal{E}_j(x). \tag{5.2}$$

- 8: Partition the particles, i.e., the set $\{1, \dots, N\}$, into disjoint sets $\mathcal{P}_k^1, \dots, \mathcal{P}_k^{N/n_N}$ of size n_N .
- 9: **for** $n = 1, \dots, N/n_N$
- 10: Compute the consensus point $y_{\alpha}(\widehat{\rho}_{Y,k\Delta t}^{N/n_N})$ according to Equation (1.4) with objective \mathcal{E}_{batch} from the particles in \mathcal{P}_k^n , i.e., with the empirical measure $\widehat{\rho}_{Y,k\Delta t}^{N/n_N} = \frac{1}{n_N} \sum_{i \in \mathcal{P}_k^n} \delta_{Y_{k\Delta t}^i}$.
- 11: Update either all particles (full update) or only the particles in the current batch \mathcal{P}_k^n (partial update) according to a discretized version of the PSO dynamics (1.2).
- 12: **if** $k > 0$ and $|y_{\alpha}(\rho_{Y,k\Delta t}^N) - y_{\alpha}(\rho_{Y,(k-1)\Delta t}^N)|$ is too small despite stopping criterion not fulfilled
- 13: Perform an independent Brownian motion for the positions or velocities of all particles.
- 14: **end if**
- 15: Set $k = k + 1$.
- 16: **end for**
- 17: **end for**
- 18: Check the stopping criterion and **break** if fulfilled. If not, employ the optional strategies described at the end of Section 5.1, set $epoch = epoch + 1$ and continue.
- 19: **end while**
- 20: Compute the consensus point $y_{\alpha}(\widehat{\rho}_{Y,T}^N)$ according to Equation (1.4) with objective \mathcal{E} from all particles, i.e., with $\widehat{\rho}_{Y,T}^N = \frac{1}{N} \sum_{i=1}^N \delta_{Y_T^i}$.

computationally efficient implementation of PSO. The mini-batch idea is present on two different levels. In line 7, the objective is defined with respect to a batch of the training data of size $n_{\mathcal{E}}$, meaning that only a subsample of the data is considered. One epoch is completed after each data sample was seen exactly once, i.e., after $M/n_{\mathcal{E}}$ steps. At each step the consensus point y_{α} has to be computed, for which \mathcal{E}_{batch} needs to be evaluated for N particles. This still constitutes the most significant computational effort. However, the mini-batch idea can be leveraged for a second time. In the **for** loop in line 9 we partition the particles into sets of size n_N and perform the updates of line 11 only for the n_N particles in the respective subset. Since this is embarrassingly parallel, a parallel machine can be deployed to reduce the runtime by up to a factor p (the number of available processors). While this is referred to as partial update, alternatively, on a sequential architecture, a full update can be made at every iteration, requiring all N particles to be updated in line 11. Apart from lowering the required computing

resources tremendously, these mini-batch ideas actually improve the stability of the method and the capability of finding good optima by introducing more stochasticity into the algorithm.

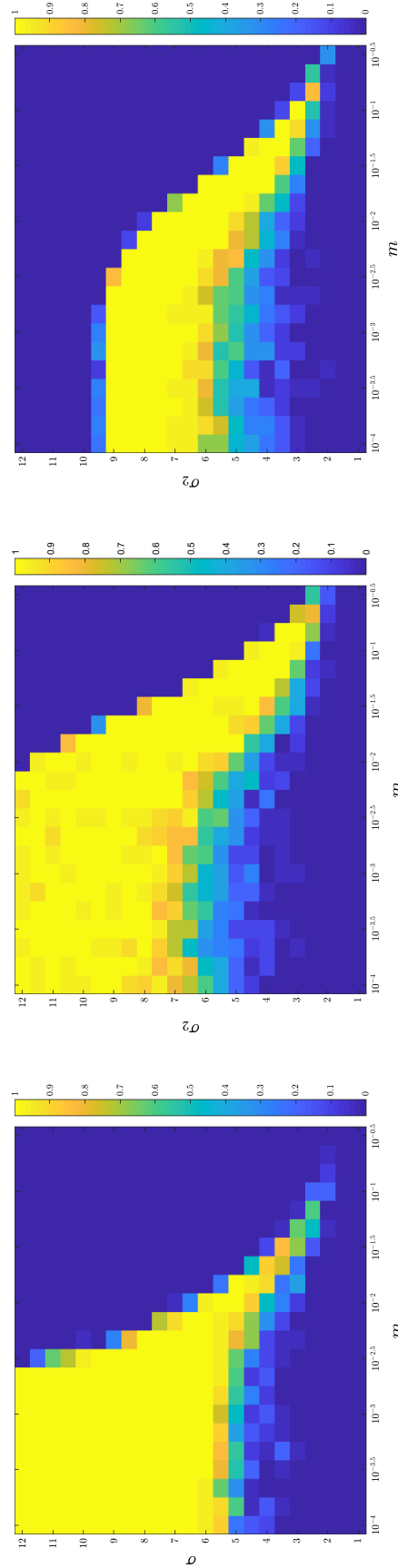
Concerning additional computational complexity due to the usage of memory effects, let us point out that, except for the required storage of the local (historical) best positions and their objective values, the update rule (1.3) in combination with the partial update allows to include such mechanisms with no additional cost by keeping track of the objective values of the local best positions. In such case, only one function call of each \mathcal{E}_{batch} per epoch and per particle is necessary, which is optimal and coincides with PSO without memory effects or CBO. A different realization of (1.2b) might result in a higher cost.

Implementational aspects: A discretization of the SDE (1.2) in line 11 can be obtained for instance from a simple Euler-Maruyama or semi-implicit scheme [23, 43], see, e.g., [21, Equation (6.3)]. In our numerical experiments below Equation (1.3) is used for updating the local best position, which corresponds to $\kappa = 1/(2\Delta t)$, $\theta = 0$, and $\beta = \infty$. Furthermore, the friction parameter is set according to $\gamma = 1 - m$, which is a typical choice in the literature. Let us also remark that a numerically stable computation of the consensus point in lines 10 and 20 for $\alpha \gg 1$ can be obtained by replacing \mathcal{E}_{batch} with $\mathcal{E}_{batch} - \tilde{\mathcal{E}}$, where $\tilde{\mathcal{E}} := \min_{i \in \mathcal{P}_k^n} \mathcal{E}_{batch}(Y_{k\Delta t}^i)$.

Cooling and evolutionary strategies: The PSO algorithm can be divided into two phases, an exploration phase, where the energy landscape is searched coarsely, and a determination phase, where the final output is identified. While the former benefits from small α and large diffusion parameters, in the latter, $\alpha \gg 1$ guarantees the selection of the best solution. A cooling strategy inspired from simulated annealing allows to start with moderate α and relatively large diffusion parameters σ_1, σ_2 . After each epoch, α is multiplied by 2, while the diffusion parameters follow the schedule $\sigma = \sigma / \log(epoch + 2)$ for $\sigma \in \{\sigma_1, \sigma_2\}$. Such strategy was proposed in [9, Section 4] for CBO. In order to further reduce computational complexity, the provable decay of the variance suggests to decrease the number of agents by discarding particles in accordance with the empirical variance. A possible schedule for the number of agents proposed in [20, Section 2.2] is to set $N_{epoch+1} = \lceil N_{epoch}((1 - \mu) + \mu \tilde{\Sigma}_{epoch} / \Sigma_{epoch}) \rceil$ for $\mu \in [0, 1]$ and where Σ_{epoch} and $\tilde{\Sigma}_{epoch}$ denote the empirical variances of the N_{epoch} particles at the beginning and at the end of the current epoch.

5.2 Numerical Experiments for the Rastrigin Function

Before turning to high-dimensional optimization problems, let us discuss the parameter choices of PSO in moderate dimensions ($d = 20$) at the example of the well-known Rastrigin benchmark function $\mathcal{E}(v) = \sum_{k=1}^d v_k^2 + \frac{5}{2}(1 - \cos(2\pi v_k))$, which meets the requirements of Assumption 1 despite being highly non-convex with many spurious local optima. To narrow down the number of tunable parameters, we let $\gamma = 1 - m$, choose $\alpha = 100$, $N = 100$, and update the local best position (if present) according to Equation (1.3), i.e., $\kappa = 1/(2\Delta t)$, $\theta = 0$, and $\beta = \infty$. We moreover let $\lambda_2 =$



(a) PSO without memory

(b) PSO with memory but no local best drift ($\lambda_1 = 0, \sigma_1 = 0$)

(c) PSO with memory and local best drift ($\lambda_1 = 0.4, \sigma_1 = 0.4\sigma_2$)

Fig. 2 Phase transition diagrams comparing PSO without and with memory effects for different inertia parameters m and noise coefficients σ (PSO without memory) and σ_2 (PSO with memory). The empirical success probability is computed from 25 runs and depicted by color from zero (blue) to one (yellow)

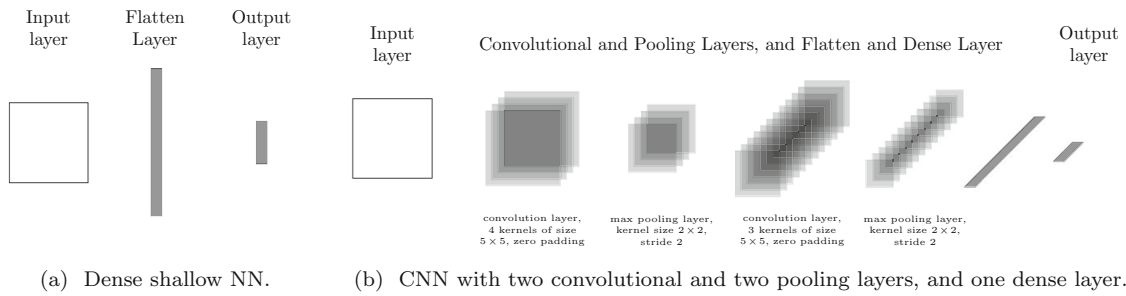


Fig. 3 Architectures of the NNs used in the experiments of Sect. 5.3, cf. [19, Section 4]

1 (or $\lambda = 1$ for PSO without memory) and $\Delta t = 0.01$, which are such that the algorithm either finds consensus or explodes within the time horizon $T = 100$ in all instances. For simplicity we assume that $\sigma_1 = \lambda_1 \sigma_2$. The algorithm is initialized with positions distributed according to $\mathcal{N}((2, \dots, 2), 4\text{Id})$ and velocities according to $\mathcal{N}((0, \dots, 0), \text{Id})$. In Fig. 2 we depict the phase diagram comparing the success probability of PSO for different parameter choices of the inertia parameter m and the diffusion parameter σ or σ_2 , respectively.

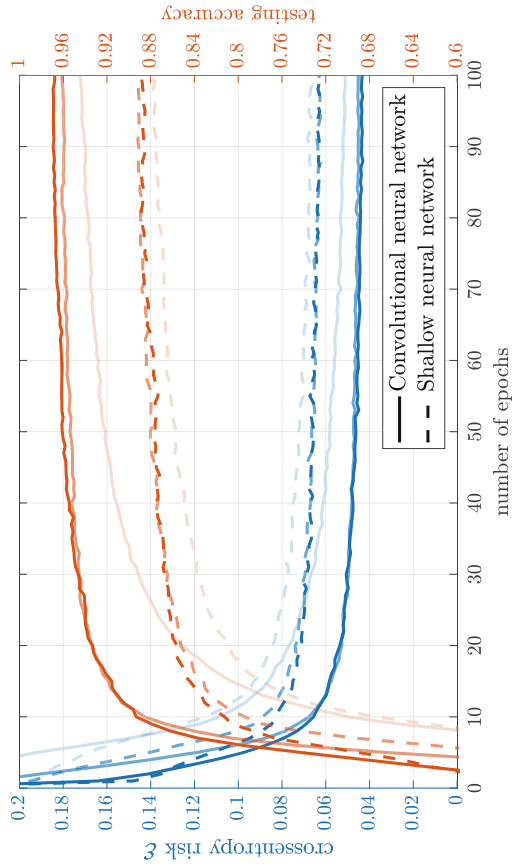
We observe that for m fixed there is a noise threshold above which the dynamics explodes. While smaller m permit a larger flexibility in the used noise, they require an individual minimal noise level. Further numerical experiments suggest however that increasing the number of particles N allows for a lower minimal noise level. There are subtle differences between PSO without and with memory, but they are not decisive as in applications also confirmed by the numerical experiments in Sect. 5.3, [22, Section 5.3] as well as the survey paper [21, Section 6.3].

5.3 A Machine Learning Application

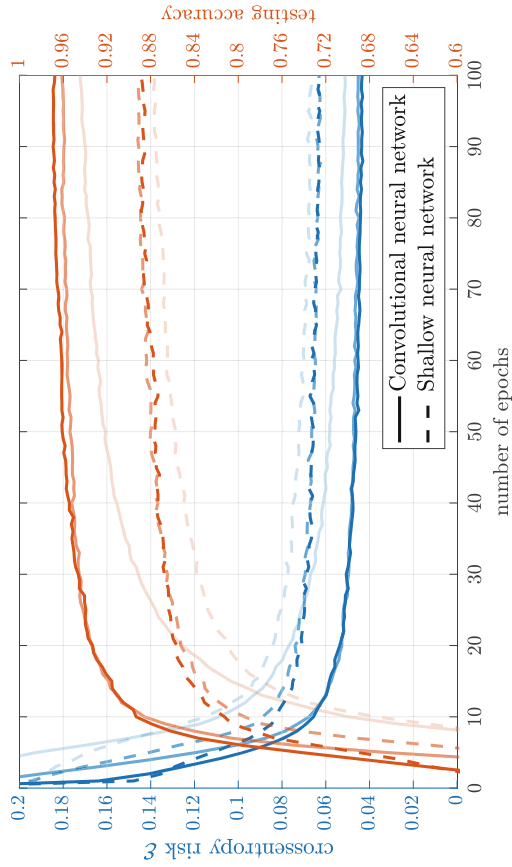
We now showcase the practicability of PSO as implemented in Algorithm 1 at the example of a very competitive high-dimensional benchmark problem in machine learning, the classification of handwritten digits. In what follows we train a shallow and a convolutional NN (CNN) classifier for the MNIST dataset [34]. Let us point out, that it is not our objective to challenge the state of the art by employing the most sophisticated model (deep CNNs achieve near-human performance of more than 99.5% accuracy). Instead, we want to demonstrate that PSO reaches results comparable to SGD with backpropagation, while at the same time relying exclusively on the evaluation of \mathcal{E} .

In our experiment we use NNs with architectures as depicted in Fig. 3.

The input is a black-and-white image represented by a (28×28) -dimensional matrix with entries between 0 and 1. For the shallow NN (see Fig. 3a), the flattened image is passed through a dense layer $\text{ReLU}(W \cdot + b)$ with trainable weights $W \in \mathbb{R}^{10 \times 728}$ and bias $b \in \mathbb{R}^{10}$. Our CNN (see Fig. 3b) is similar to LeNet-1, cf. [33, Section III.C.7]. Each dense or convolution layer has a ReLU activation and is followed by a batch normalization layer to speed up the training process. Eventually, the final layers of both NNs apply a softmax activation function allowing to interpret the 10-dimensional output vector as a probability distribution over the digits.



(a) PSO for three different memory settings: without memory (lightest lines); with memory but without local best drift, i.e., $\lambda_1 = 0$, $\sigma_1 = 0$ (line with intermediate opacity); with memory with local best drift $\lambda_1 = 0.4$, $\sigma_1 = \lambda_1 \sigma_2$ (darkest lines).



(b) PSO with memory but without local best drift for three different inertia parameters $m \in \{0.1, 0.2, 0.4\}$ (increasing opacity for larger m). Note that, for reference, the lines with intermediate opacity coincide with the ones of Fig. 4a.

Fig. 4 Comparison of the performances of a shallow (dashed lines) and convolutional (solid lines) NN with architectures as described in Fig. 3, when trained with PSO as in Algorithm 1. Depicted are the accuracies on a test dataset (orange lines) and the values of the objective function \mathcal{E} (blue lines) evaluated on a random sample of the training set of size 10000

We denote by θ the trainable parameters of the NNs, which are 7850 for the shallow NN and 2112 for the CNN. They are learned by minimizing $\mathcal{E}(\theta) = \frac{1}{M} \sum_{j=1}^M \ell(f_\theta(x^j), y^j)$, where f_θ denotes the forward pass of the NN, (x^j, y^j) the j th image-label tuple and ℓ the categorical crossentropy loss $\ell(\hat{y}, y) = -\sum_{k=0}^9 y_k \log(\hat{y}_k)$. The performance is measured by counting the number of successful predictions on a test set. We use a train-test split of 60000 training and 10000 test images. For our experiments we choose $\lambda_2 = 1$, $(\sigma_2)_{initial} = \sqrt{0.4}$, $\alpha_{initial} = 50$, $\Delta t = 0.1$ and update the local best position according to Equation (1.3). We use $N = 100$ agents, which are initialized according to $\mathcal{N}((0, \dots, 0)^T, \text{Id})$ in position and velocity. The mini-batch sizes are $n_\mathcal{E} = 60$ and $n_N = 100$ (consequently a full update is performed in line 11) and a cooling strategy is used in line 18.

Figure 4a reports the performances for different memory settings and fixed $m = 0.2$, whereas Fig. 4b depicts the results for different inertia parameters m in the case of PSO with memory but no memory drift.

For the shallow NN, we obtain a test accuracy of above 89%, whereas the CNN achieves almost 97%. To put those numbers into perspective, when trained with SGD, a similar performance for the shallow NN, see [9, Figure 7], and a benchmark accuracy of 98.3% for a comparable CNN, cf. [33, Figure 9], are reached. As can be seen from Fig. 4a, the usage of the local best positions when computing the consensus point significantly improves the performance. The advantage of having an additional drift towards the local best position is less pronounced. Regarding the inertia parameter m in Fig. 4b, our numerical results suggest that larger m yield faster convergence.

6 Conclusions

In this paper we prove the convergence of PSO without and with memory effects to a global minimizer of a possibly nonconvex and nonsmooth objective function in the mean-field sense. Our analysis holds under a suitable well-preparation condition about the initialization and comprises a rich class of objectives which in particular includes functions with multiple global minimizers. For PSO without memory effects we furthermore quantify how well the mean-field dynamics approximates the interacting finite particle dynamics, which is implemented for numerical experiments. Since in particular the latter results does not suffer from the curse of dimensionality, we thereby prove that the numerical PSO method has polynomial complexity. With this we contribute to the completion of a mathematically rigorous understanding of PSO. Furthermore, we propose a computationally efficient and parallelizable implementation and showcase its practicability by training a CNN reaching a performance comparable to stochastic gradient descent.

It remains an open problem to extend the mean-field approximation result to the variant of PSO with memory effects or, alternatively, to devise an implementation of such effects compatible with the used proof technique. Moreover, we also leave a more thorough understanding of the influence of the parameters as well as the influence of memory effects for future, more experimental research.

Finally, we believe that the analysis framework of this and prior works on CBO [8, 18, 42] motivates to investigate also other renowned metaheuristic algorithms through the lens of a mean-field limit.

Funding Open Access funding enabled and organized by Projekt DEAL. H.H. is partially supported by the Pacific Institute for the Mathematical Sciences (PIMS) postdoctoral fellowship. J.Q. is partially supported by the National Science and Engineering Research Council of Canada (NSERC) and by the start-up funds from the University of Calgary. K.R. acknowledges the financial support from the Technical University of Munich – Institute for Ethics in Artificial Intelligence (IEAI). The authors gratefully acknowledge the compute and data resources provided by the Leibniz Supercomputing Centre (LRZ).

Declarations

Competing interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Aarts, E., Korst, J.: Simulated annealing and Boltzmann machines. A stochastic approach to combinatorial optimization and neural computing. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, Chichester (1989)
2. Arnold, L.: Stochastic differential equations: Theory and applications. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney (1974). Translated from the German
3. Azencott, R. (ed.): Simulated annealing: Parallelization techniques. Wiley (1992)
4. van den Bergh, F.: An analysis of particle swarm optimizers. Ph.D. thesis, University of Pretoria (2007)
5. van den Bergh, F., Engelbrecht, A.P.: A convergence proof for the particle swarm optimiser. *Fund. Inform.* **105**(4), 341–374 (2010)
6. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Stochastic mean-field limit: non-Lipschitz forces and swarming. *Math. Models Methods Appl. Sci.* **21**(11), 2179–2210 (2011)
7. Bruned, V., Mas, A., Włodarczyk, S.: Weak convergence of particle swarm optimization. arXiv preprint [arXiv:1811.04924](https://arxiv.org/abs/1811.04924) (2018)
8. Carrillo, J.A., Choi, Y.P., Totzeck, C., Tse, O.: An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.* **28**(6), 1037–1066 (2018)
9. Carrillo, J.A., Jin, S., Li, L., Zhu, Y.: A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.* **27**(suppl.), Paper No. S5, 1–22 (2021)
10. Cipriani, C., Huang, H., Qiu, J.: Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Math. Anal.* **54**(3), 3091–3121 (2022)
11. Clerc, M., Kennedy, J.: The particle swarm—explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evol. Comput.* **6**(1), 58–73 (2002). <https://doi.org/10.1109/4235.985692>
12. Dembo, A., Zeitouni, O.: Large deviations techniques and applications, Applications of Mathematics (New York), vol. 38, 2nd edn. Springer, New York (1998)
13. Ding, Z., Chen, S., Li, Q., Wright, S.: On the global convergence of gradient descent for multi-layer resnets in the mean-field regime. arXiv preprint [arXiv:2110.02926](https://arxiv.org/abs/2110.02926) (2021)

14. Dréo, J., Pétrowski, A., Siarry, P., Taillard, E.: Metaheuristics for hard optimization. Springer, Berlin (2006). Methods and case studies, Translated from the 2003 French original by Amitava Chatterjee
15. Fogel, D.B.: Evolutionary Computation. Toward a New Philosophy of Machine Intelligence, 2nd edn. IEEE Press, Piscataway (2000)
16. Fornasier, M., Huang, H., Pareschi, L., Sünnen, P.: Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.* **30**(14), 2725–2751 (2020)
17. Fornasier, M., Huang, H., Pareschi, L., Sünnen, P.: Anisotropic diffusion in consensus-based optimization on the sphere. *SIAM J. Optim.* **32**(3), 1984–2012 (2022)
18. Fornasier, M., Klock, T., Riedl, K.: Consensus-based optimization methods converge globally. arXiv preprint [arXiv:2103.15130](https://arxiv.org/abs/2103.15130) (2021)
19. Fornasier, M., Klock, T., Riedl, K.: Convergence of anisotropic consensus-based optimization in mean-field law. In: J.L. Jiménez Laredo, J.I. Hidalgo, K.O. Babaagba (eds.) *Applications of Evolutionary Computation*, pp. 738–754. Springer, Cham (2022)
20. Fornasier, M., Pareschi, L., Huang, H., Sünnen, P.: Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.* **22**(237), 1–55 (2021)
21. Grassi, S., Huang, H., Pareschi, L., Qiu, J.: Mean-field particle swarm optimization. arXiv preprint [arXiv:2108.00393](https://arxiv.org/abs/2108.00393) (2021)
22. Grassi, S., Pareschi, L.: From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Math. Models Methods Appl. Sci.* **31**(8), 1625–1657 (2021)
23. Higham, D.J.: An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.* **43**(3), 525–546 (2001)
24. Holland, J.H.: *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence.* University of Michigan Press, Ann Arbor, Mich. (1975)
25. Huang, H., Liu, J.G., Pickl, P.: On the mean-field limit for the Vlasov-Poisson-Fokker-Planck system. *J. Stat. Phys.* **181**(5), 1915–1965 (2020)
26. Huang, H., Qiu, J.: On the mean-field limit for the consensus-based optimization. *Mathematical Methods in the Applied Sciences*, pp. 1–18 (2022)
27. Jabin, P.E., Wang, Z.: Mean field limit for stochastic particle systems. In: *Active particles. Vol. 1. Advances in theory, models, and applications, Model. Simul. Sci. Eng. Technol.*, pp. 379–402. Birkhäuser/Springer, Cham (2017)
28. Jin, S., Li, L., Liu, J.G.: Random batch methods (RBM) for interacting particle systems. *J. Comput. Phys.* **400**(108877), 1–30 (2020)
29. Kadanoff, L.P.: More is the same; phase transitions and mean field theories. *J. Stat. Phys.* **137**(5–6), 777–797 (2009)
30. Kennedy, J.: The particle swarm: social adaptation of knowledge. In: *Proceedings of 1997 IEEE International Conference on Evolutionary Computation*, pp. 303–308. IEEE (1997). [10.1109/ICEC.1997.592326](https://doi.org/10.1109/ICEC.1997.592326)
31. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948. IEEE (1995). [10.1109/ICNN.1995.488968](https://doi.org/10.1109/ICNN.1995.488968)
32. Kushner, H., Yin, G.G.: *Stochastic approximation and recursive algorithms and applications*, vol. 35. Springer Science & Business Media (2003)
33. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
34. LeCun, Y., Cortes, C., Burges, C.: MNIST handwritten digit database (2010). <http://yann.lecun.com/exdb/mnist/>
35. Lin, S.W., Ying, K.C., Chen, S.C., Lee, Z.J.: Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Syst. Appl.* **35**(4), 1817–1824 (2008)
36. Mei, S., Montanari, A., Nguyen, P.M.: A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **115**(33), E7665–E7671 (2018)
37. Miclo, L.: Recuit simulé sur \mathbb{R}^n . étude de l'évolution de l'énergie libre. In: *Annales de l'IHP Probabilités et statistiques*, vol. 28, pp. 235–266 (1992)
38. Miller, P.D.: *Applied Asymptotic Analysis, Graduate Studies in Mathematics*, vol. 75. American Mathematical Society, Providence, RI (2006)
39. Øksendal, B.: *Stochastic Differential Equations: An Introduction with Applications*, 6th edn. Springer, Berlin (2003)
40. Özcan, E., Mohan, C.K.: Analysis of a simple particle swarm optimization system (1998)

41. Panigrahi, B.K., Shi, Y., Lim, M.H.: Handbook of swarm intelligence: concepts, principles and applications, vol. 8. Springer Science & Business Media (2011)
42. Pinnau, R., Totzeck, C., Tse, O., Martin, S.: A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.* **27**(1), 183–204 (2017)
43. Platen, E.: An introduction to numerical methods for stochastic differential equations. In: *Acta numerica*, *Acta Numer.*, vol. 8, pp. 197–246. Cambridge University Press, Cambridge (1999)
44. Poli, R.: Mean and variance of the sampling distribution of particle swarm optimizers during stagnation. *IEEE Trans. Evol. Comput.* **13**(4), 712–721 (2009)
45. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. *Swarm Intell* **1**(1), 33–57 (2007). <https://doi.org/10.1007/s11721-007-0002-0>
46. Rastrigin, L.A.: The convergence of the random search method in the external control of many-parameter system. *Autom. Remote Control* **24**, 1337–1342 (1963)
47. Revuz, D., Yor, M.: Continuous martingales and Brownian motion, *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, vol. 293, third edn. Springer-Verlag, Berlin (1999)
48. Riedl, K.: Leveraging memory effects and gradient information in consensus-based optimization: On global convergence in mean-field law. [arXiv:2211.12184](https://arxiv.org/abs/2211.12184) (2022)
49. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* **22**, 400–407 (1951)
50. Royer, G.: An initiation to logarithmic Sobolev inequalities. 5. American Mathematical Soc. (2007)
51. Schmitt, M., Wanka, R.: Particle swarm optimization almost surely finds local optima. *Theor. Comput. Sci.* **561**(Part A), 57–72 (2015)
52. Sznitman, A.S.: Topics in propagation of chaos. In: *École d’Été de Probabilités de Saint-Flour XIX – 1989*, *Lecture Notes in Math.*, vol. 1464, pp. 165–251. Springer, Berlin (1991)
53. Tang, W., Zhou, X.Y.: Tail probability estimates of continuous-time simulated annealing processes. *Numerical Algebra, Control and Optimization* (2022)
54. Totzeck, C., Wolfram, M.T.: Consensus-based global optimization with personal best. *Math. Biosci. Eng.* **17**(5), 6026–6044 (2020)
55. Witt, C.: Theory of particle swarm optimization. In: *Theory of randomized search heuristics*, Ser. *Theor. Comput. Sci.*, vol. 1, pp. 197–223. World Sci. Publ., Hackensack, NJ (2011)
56. Yuan, Q., Yin, G.: Analyzing convergence and rates of convergence of particle swarm optimization algorithms using stochastic approximation methods. *IEEE Trans. Autom. Control* **60**(7), 1760–1773 (2014)
57. Zhang, Y., Wang, S., Ji, G.: A comprehensive survey on particle swarm optimization algorithm and its applications. *Math. Probl. Eng.* **931256**, 1–38 (2015)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

License for [PSO].

The permission to reprint and include the material is printed on the next page(s).



RightsLink

On the Global Convergence of Particle Swarm Optimization Methods

SPRINGER NATURE**Author:** Hui Huang et al**Publication:** Applied Mathematics and Optimization**Publisher:** Springer Nature**Date:** May 31, 2023*Copyright © 2023, The Author(s)*

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Springer Nature](#)

© 2024 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Data Security and Privacy](#)
| [For California Residents](#) | [Terms and Conditions](#) Comments? We would like to hear from you. E-mail us at customer@copyright.com

Paper P8

Consensus-Based Optimization for Saddle Point Problems

H. Huang, J. Qiu, and K. Riedl
SIAM J. Control Optim. (2024)

Paper Summary of [CBO-SP]⁴¹

In the paper “Consensus-Based Optimization for Saddle Point Problems,” published in the *SIAM Journal on Control and Optimization*, we propose and investigate the CBO-SP method (6.4) from both an analytical point of view by taking a mean-field perspective and experimentally.

CBO-SP is a novel multi-particle metaheuristic derivative-free optimization method for saddle point problems capable of provably finding global Nash equilibria, i.e., solving problems of the form (6.1). It is substantially inspired by its optimization relative CBO [Pin+17].

Following the idea of swarm intelligence, our CBO-SP algorithm [CBO-SP, Algorithm 1] employs two groups of interacting particles, one of which performs a minimization over one variable while the other performs a maximization over the other variable. The two groups constantly exchange information through a suitably weighted average, the consensus point. In [CBO-SP, Section 2.1], we prove the well-posedness of the continuous-time analog of the CBO-SP algorithm, which is described through a system of SDEs. This paradigm permits a passage to the mean-field limit (whose well-posedness is shown in [CBO-SP, Section 2.2]), which makes the method amenable to a theoretical convergence analysis. In particular, under reasonable assumptions on the objective function [CBO-SP, Definition 9], which most notably include nonconvex-nonconcave objectives, and reasonable assumptions about the initialization [CBO-SP, Definition 10], rigorous convergence guarantees [CBO-SP, Theorem 11] can be obtained. Following the analytical framework of [Car+18; Car+21], which was developed for CBO in the optimization setting, we first prove, under certain well-preparedness conditions, consensus formation of the mean-field dynamics at some location. In a second and consecutive step, which involves suitable choices of the parameters of the method, this consensus is shown to have properties that are typical for saddle points. Eventually, under a suitable condition on the objective function, the aforementioned properties imply that the found consensus is close to a saddle point. Our numerical investigations, which comprise illustrative numerical experiments [CBO-SP, Section 5.2] as well as the task of solving a quadratic game [CBO-SP, Section 5.3], provide numerical evidence for the success and efficiency of the proposed CBO-SP algorithm.

KR’s Contributions. JQ suggested to extend the idea of CBO to saddle point problems. Together with JQ, HH and KR devised a suitable numerical scheme for finding global Nash equilibria giving rise to CBO-SP, for which KR proved the well-posedness of the interacting particle system and of the mean-field limit dynamics. The convergence properties of the mean-field limit under certain well-preparedness assumptions of the initial data and the parameters of the scheme were then analyzed by HH and KR. KR conducted the numerical experiments and wrote large parts of the paper, which was proofread and refined by JQ and HH.

⁴¹In this section, we follow [CBO-SP, Abstract].

The following document is a reprint of

[CBO-SP] H. Huang, J. Qiu, and K. Riedl. “Consensus-Based Optimization for Saddle Point Problems.” In: *SIAM J. Control Optim.* 62.2 (2024), pp. 1093–1121.

The permission to reprint and include the material is provided after the reprint.

CONSENSUS-BASED OPTIMIZATION FOR SADDLE POINT PROBLEMS*

HUI HUANG[†], JINNIAO QIU[‡], AND KONSTANTIN RIEDL^{§¶}

Abstract. In this paper, we propose consensus-based optimization for saddle point problems (CBO-SP), a novel multi-particle metaheuristic derivative-free optimization method capable of provably finding global Nash equilibria. Following the idea of swarm intelligence, the method employs two groups of interacting particles, one which performs a minimization over one variable while the other performs a maximization over the other variable. The two groups constantly exchange information through a suitably weighted average. This paradigm permits a passage to the mean-field limit, which makes the method amenable to theoretical analysis, and it allows to obtain rigorous convergence guarantees under reasonable assumptions about the initialization and the objective function, which most notably include nonconvex-nonconcave objectives. We further provide numerical evidence for the success of the algorithm.

Key words. saddle point problems, Nash equilibria, nonconvex-nonconcave, derivative-free optimization, metaheuristics, consensus-based optimization, Fokker–Planck equations

MSC codes. 90C47, 65C35, 65K05, 90C56, 35Q90, 35Q83

DOI. 10.1137/22M1543367

1. Introduction. Optimization problems where the goal is to find the best possible objective value for the worst-case scenario can be formulated as minimax optimization problems of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{E}(x, y).$$

To be more specific, given a class of objective functions $\{\mathcal{E}(\cdot, y), y \in \mathcal{Y}\}$, the aim is to determine the argument $x^* \in \mathcal{X}$ that leads to the smallest objective value even for the worst-case function parametrized by $y^* \in \mathcal{Y}$. Such type of problems were originally formulated in two-player zero-sum game theory [47] but now arise in many areas in mathematics, biology, the social sciences, and especially economics [34]. Diverse applications may be found in engineering, operational research, biology, ecology, finance, economics, energy industry, environmental sciences, and so on. In the last few years, minimax optimization has also experienced substantial attention from the signal processing community, due to its connection to distributed processing [8], robust transceiver design [26], and communication in the presence of jammers [16]. Moreover,

*Received by the editors December 25, 2022; accepted for publication (in revised form) September 22, 2023; published electronically March 25, 2024.

<https://doi.org/10.1137/22M1543367>

Funding: The second author's research was partially supported by the National Science and Engineering Research Council of Canada (NSERC) and by the PIMS-Europe Fellowship. The third author's research was supported by the Institute for Ethics in Artificial Intelligence (IEAI), Technical University of Munich.

[†]Department of Mathematics and Scientific Computing, University of Graz, 8010 Graz, Austria (hui.huang@uni-graz.at).

[‡]Department of Mathematics and Statistics, University of Calgary, T2N 1N4 Calgary, Canada (jinniao.qiu@ucalgary.ca).

[§]Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany (konstantin.riedl@ma.tum.de).

[¶]Munich Center for Machine Learning, 80593 Munich, Germany.

in modern machine learning, several problems are formulated as minimax optimization, such as the training of generative adversarial networks (GANs) [17], multi-agent reinforcement learning [38], fair machine learning [29], and adversarial training [30]. For example, when training GANs, x models the parameter of a generator, usually a neural network, whose aim is to generate synthetic data with the same statistics as a given training set, while y represents the parameters of a competing discriminator, who has to distinguish generated data by the generator from data of the true distribution. Relatedly, in adversarial machine learning, one aims at learning the parameters x of a model in a robust manner by exposing it during training to possible adversarial attacks modeled by y . Both examples can be interpreted as a game between two neural networks trained in an adversarial manner until some kind of equilibrium is reached.

In a two-player zero-sum game, the joint payoff function $\mathcal{E}(x, y)$ encodes the gain of the maximization player whose action is to choose $y \in \mathcal{Y}$, as well as the loss of the minimization player controlling the action $x \in \mathcal{X}$. In simultaneous games, each player chooses its action without the knowledge of the action chosen by the other player, so both players act simultaneously. Conversely, in sequential games there is an intrinsic order according to which the players take their actions, meaning that the ordering of the minimization and maximization matters, i.e., it plays a priori a role whether $\min_x \max_y$ or $\max_y \min_x$. GANs and adversarial training, for instance, are in fact sequential games in their standard formulations. In the classical case, where the payoff function \mathcal{E} is *convex-concave* (i.e., $\mathcal{E}(\cdot, y)$ is convex for all $y \in \mathcal{Y}$ and $\mathcal{E}(x, \cdot)$ is concave for all $x \in \mathcal{X}$), the intrinsic order of sequential games does not matter under an additional compactness assumption on either \mathcal{X} or \mathcal{Y} by the well-known minimax theorems of Sion and von Neumann (see [45, 46]). However, nowadays, most modern applications in signal processing and machine learning entail the setting of *nonconvex-nonconcave* minimax problems, where the minimization and maximization problems are potentially nonconvex and nonconcave. This is significantly more complicated, and available tool sets and theories are very limited; see the review paper [42].

A well-known notion of optimality originating from game theory is the one of Nash equilibria (also referred to as saddle points) [35], where neither of the players has anything to gain by changing only his own strategy. This concept is formalized within the following definition.

DEFINITION 1. *A point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called a Nash equilibrium or saddle point of a function \mathcal{E} if it holds that*

$$\mathcal{E}(x^*, y) \leq \mathcal{E}(x^*, y^*) \leq \mathcal{E}(x, y^*) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

or, equivalently, if

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{E}(x, y) = \mathcal{E}(x^*, y^*) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \mathcal{E}(x, y).$$

To keep the notation concise, we write \mathcal{E}^* for $\mathcal{E}(x^*, y^*)$ in what follows.

In the *convex-concave* setting, an approximate Nash equilibrium can be found efficiently by variants of gradient descent-ascent (GDA) algorithms [4, 18], which alternate between one or more gradient descent steps in the x -variable and gradient ascent steps in the y -coordinate. Indeed, even if $\mathcal{E}(x, y)$ is either concave in y or convex in x , there are some multistep GDA algorithms available; see [36, 42], for instance. However, as soon as the payoff function becomes *nonconvex-nonconcave*, finding a global equilibrium is in general an NP-hard problem [33]. For this reason,

some recent works, such as [9, 31], consider a local version of equilibria. More precisely, a point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is called a local Nash equilibrium if there exists some $\delta > 0$ such that (x^*, y^*) satisfies Definition 1 in a δ -neighborhood of (x^*, y^*) . Local Nash equilibria can be characterized in terms of the so-called quasi-Nash equilibrium condition [39] or the first-order Nash equilibrium condition [36]. Even so, we mention two recent works where special classes of nonconvex-nonconcave payoff functions are considered. When $\mathcal{E}(x, y)$ is weakly convex in x and weakly concave in y and the associated Minty variational inequality admits a solution, Liu et al. [25] employ the inexact proximal point method and prove the first-order convergence, while under the so-called sufficiently bilinear condition, the stochastic Hamiltonian method is investigated by Loizou et al. [28]. In this work, we shall drop such restrictions and the gradient dependence in the algorithms and consider a *zero-order* (derivative-free) method with rigorous convergence guarantees. Note that the family of population-based algorithms, such as Particle Swarm Optimization (PSO) [22], has been adapted to solve min-max problems as done, for instance, in [23, 24, 44]. One straightforward approach is to treat the min-max problem as a minimization problem and embed the maximization part in the calculation of the objective values [24]. Alternatively, a multi-PSO strategy [23, 44] may be employed, where the min-max problem is converted into two optimization problems, one being a maximization problem and the other a minimization problem. Two PSO algorithms are then used to solve these two optimization problems, respectively, and they are run independently. Each PSO is treated as a changing environment of the other PSO, allowing them to cooperate through the calculation of the objective. Both approaches cannot avoid the necessity of nested loops/circles of optimization algorithms, which significantly increases the time complexity.

In the present paper, we propose a zero-order consensus-based optimization method for finding the global Nash equilibrium (x^*, y^*) of a smooth objective function $\mathcal{E} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ with $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$, which is designed to be amenable to a rigorous theoretical convergence analysis, missing so far in the literature on population-based methods for min-max problems. The dynamics of the algorithm is inspired by consensus-based optimization, a paradigm for global nonconvex minimizations, which was introduced by the authors of [40]. Their method employs a system of interacting particles which explore the energy landscape in order to form a global consensus about the global minimizer of the objective function as time passes. Taking inspiration from this concept, let us consider two sets of particles $(X^i)_{i=1}^{N_1}$ and $(Y^i)_{i=1}^{N_2}$ of potentially different size, one for minimization and the other for maximization. Each individual particle of either set is formally described by a stochastic process. In order to achieve consensus about the equilibrium point of \mathcal{E} , the particles interact through a system of stochastic differential equations (SDEs) of the form

(1a)

$$dX_t^i = -\lambda_1 \left(X_t^i - x_\alpha^Y(\hat{\rho}_{X,t}^{N_1}) \right) dt + \sigma_1 D \left(X_t^i - x_\alpha^Y(\hat{\rho}_{X,t}^{N_1}) \right) dB_t^{X,i}, \quad \hat{\rho}_{X,t}^{N_1} = \frac{1}{N_1} \sum_{i=1}^{N_1} \delta_{X_t^i},$$

(1b)

$$dY_t^i = -\lambda_2 \left(Y_t^i - y_\beta^X(\hat{\rho}_{Y,t}^{N_2}) \right) dt + \sigma_2 D \left(Y_t^i - y_\beta^X(\hat{\rho}_{Y,t}^{N_2}) \right) dB_t^{Y,i}, \quad \hat{\rho}_{Y,t}^{N_2} = \frac{1}{N_2} \sum_{i=1}^{N_2} \delta_{Y_t^i},$$

which is complemented by suitable initial conditions $X_0^i \sim \rho_{X,0} \in \mathcal{P}(\mathbb{R}^{d_1})$ for $i = 1, \dots, N_1$ and $Y_0^i \sim \rho_{Y,0} \in \mathcal{P}(\mathbb{R}^{d_2})$ for $i = 1, \dots, N_2$ and where $((B_t^{X,i})_{t \geq 0})_{i=1, \dots, N_1}$

and $((B_t^{Y,i})_{t \geq 0})_{i=1, \dots, N_2}$ are independent standard Brownian motions in \mathbb{R}^{d_1} and \mathbb{R}^{d_2} , respectively. Moreover, $\widehat{\rho}_{X,t}^{N_1}$ and $\widehat{\rho}_{Y,t}^{N_2}$ denote the empirical measures of the particles' x - and y -positions, respectively. While the dynamics (1a) performs minimization in the x -variable, (1b) performs maximization in the y -coordinate. This is encoded in the computation of the so-called consensus point $(x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}), y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}))$, whose components are given by

(2a)

$$x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}) = \int x \frac{\omega_\alpha(x, \int y d\widehat{\rho}_{Y,t}^{N_2}(y))}{\|\omega_\alpha(\cdot, \int y d\widehat{\rho}_{Y,t}^{N_2}(y))\|_{L_1(\widehat{\rho}_{X,t}^{N_1})}} d\widehat{\rho}_{X,t}^{N_1}(x) \quad \text{with } \omega_\alpha(x, y) := \exp(-\alpha \mathcal{E}(x, y)),$$

(2b)

$$y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}) = \int y \frac{\omega_{-\beta}(\int x d\widehat{\rho}_{X,t}^{N_1}(x), y)}{\|\omega_{-\beta}(\int x d\widehat{\rho}_{X,t}^{N_1}(x), \cdot)\|_{L_1(\widehat{\rho}_{Y,t}^{N_2})}} d\widehat{\rho}_{Y,t}^{N_2}(y) \quad \text{with } \omega_{-\beta}(x, y) := \exp(\beta \mathcal{E}(x, y)).$$

Attributed to the Laplace principle [32], $x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})$ can be interpreted as an approximation of $\arg \min_{i=1, \dots, N_1} \mathcal{E}(X_t^i, \int y d\widehat{\rho}_{Y,t}^{N_2}(y))$ as $\alpha \rightarrow \infty$, while $y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}) \approx \arg \max_{i=1, \dots, N_2} \mathcal{E}(\int x d\widehat{\rho}_{X,t}^{N_1}(x), Y_t^i)$ as $\beta \rightarrow \infty$; see, e.g., [14, equation (7)]. The dynamics of each of the particles in (1) is governed by two terms. A drift term drags the particles towards the respective component of the instantaneous consensus point $(x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}), y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}))$ and thereby expectedly improves the position of the particles. The second term injects stochasticity into the dynamics by diffusing the particles according to a scaled Brownian motion, which features the exploration of the landscape of the objective. In what follows, we use anisotropic noise, i.e., $D(\cdot) = \text{diag}(\cdot)$, which is typically more competitive in high dimensions compared to isotropic noise $D(\cdot) = \|\cdot\|_2$; see, e.g., [6, 15]. The theoretical results of this paper, however, can be obtained mutatis mutandis also in the isotropic setting.

An implementable scheme for a numerical algorithm can be obtained from (1) by a simple Euler–Maruyama time discretization [19, 41]. For details about the implementation, we refer the reader to Algorithm 1 in section 5.1.

Remark 2. While the definition of the consensus point in (2) is a natural option, there are two equally reasonable alternatives. The first possibility is to replace the mean $\int y d\widehat{\rho}_{Y,t}^{N_2}(y)$ in (2a) simply by y and integrate w.r.t. the joint measure $\widehat{\rho}_t^N$. This case would require $N_1 = N_2$. Analogously, $\int x d\widehat{\rho}_{X,t}^{N_1}(x)$ is substituted by x in (2b). The second option is to use the other component of the consensus point instead of the respective mean; i.e., $y_\beta^X(\widehat{\rho}_{Y,t}^{N_2})$ replaces $\int y d\widehat{\rho}_{Y,t}^{N_2}(y)$ in (2a) and $x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1})$ substitutes $\int x d\widehat{\rho}_{X,t}^{N_1}(x)$ in (2b).

The main motivation for using the variant as in (2) is of a theoretical nature. Using either of the other two alternatives significantly complicates the convergence analysis in sections 3 and 4.

Understanding the convergence properties of the dynamics (1) can take place either by investigating the long time behavior of the interacting particle system itself or by analyzing the macroscopic behavior of the agent density associated with (1) through a mean-field limit. This theoretical approach proved successful in [2, 3, 5, 6, 7, 12, 13, 14, 15, 43] for proving global convergence for several variants of consensus-based optimization in the setting of minimization. It is, moreover, theoretically justified by the mean-field approximation which shows that $(\widehat{\rho}_{X,t}^{N_1}, \widehat{\rho}_{Y,t}^{N_2})$ converges in some sense to a mean field law $(\rho_{X,t}, \rho_{Y,t})$ as $N_1, N_2 \rightarrow \infty$. Again, for consensus-based optimization there exist by now several results in this direction, such as [11, 14, 20],

which may be extended to CBO-SP in an immediate manner. In the setting of saddle point problems, the mean-field dynamics associated with (1) can be described by the self-consistent monoparticle dynamics

$$(3a) \quad d\bar{X}_t = -\lambda_1 (\bar{X}_t - x_\alpha^Y(\rho_{X,t})) dt + \sigma_1 D(\bar{X}_t - x_\alpha^Y(\rho_{X,t})) dB_t^X, \quad \rho_{X,t} = \int d\rho_t(\cdot, y),$$

$$(3b) \quad d\bar{Y}_t = -\lambda_2 (\bar{Y}_t - y_\beta^X(\rho_{Y,t})) dt + \sigma_2 D(\bar{Y}_t - y_\beta^X(\rho_{Y,t})) dB_t^Y, \quad \rho_{Y,t} = \int d\rho_t(x, \cdot),$$

where $\rho_t = \rho(t) = \text{Law}((\bar{X}_t, \bar{Y}_t))$ with marginals $\rho_{X,t}$ and $\rho_{Y,t}$, respectively. In particular, the measure $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^{d_1+d_2}))$ weakly satisfies the nonlinear nonlocal Fokker-Planck equation

$$(4) \quad \begin{aligned} \partial_t \rho_t &= \lambda_1 \text{div}_x((x - x_\alpha^Y(\rho_t^X)) \rho_t) + \lambda_2 \text{div}_y((y - y_\beta^X(\rho_t^Y)) \rho_t) \\ &+ \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \partial_{x_k x_k}^2 ((x - x_\alpha^Y(\rho_t^X))_k^2 \rho_t) + \frac{\sigma_2^2}{2} \sum_{k=1}^{d_2} \partial_{y_k y_k}^2 ((y - y_\beta^X(\rho_t^Y))_k^2 \rho_t). \end{aligned}$$

Contributions. Motivated by the fundamental importance of *nonconvex-nonconcave* saddle point problems in various applicational areas and the desire for having numerical algorithms with rigorous global convergence guarantees, we theoretically analyze in this work a novel consensus-based optimization method (CBO-SP) capable of tackling saddle point problems. Using mean-field analysis techniques, we rigorously prove that CBO-SP converges to saddle points as the number of interacting particles goes to infinity. Our results hold under reasonable assumptions about the objective function and under certain conditions of the well-preparation of the hyperparameters and the initial data.

1.1. Organization. In section 2, we first investigate the well-posedness of both the interacting particle system (1) of CBO-SP and its associated mean-field dynamics (3). Section 3 then presents and discusses the main theoretical statement of this work concerned with the convergence of the mean-field dynamics (3) towards saddle points of the objective function \mathcal{E} , which are proven in section 4. Section 5 contains details about the implementation of the numerical algorithm as well as instructive numerical examples which illustrate how CBO-SP works, before we conclude the paper in section 6. In the GitHub repository <https://github.com/KonstantinRiedl/CBOSaddlePoints>, we provide the MATLAB code implementing CBO-SP.

2. Well-posedness of CBO-SP and its mean-field dynamics. In the first part of this section, we provide a well-posedness result about the interacting particle system (1) of CBO-SP; i.e., we show that a process obeying (1) exists and is unique. Afterwards, we also prove the well-posedness of the nonlinear macroscopic SDE (3).

2.1. Well-posedness of the interacting particle system. To keep the notation concise in what follows, let us denote the state vector of the entire particle system (1) by $\mathbf{Z} \in \mathcal{C}([0, \infty), \mathbb{R}^{N_1 d_1 + N_2 d_2})$ with $\mathbf{Z}(t) = \mathbf{Z}_t = \left((X_t^1)^T, \dots, (X_t^{N_1})^T, (Y_t^1)^T, \dots, (Y_t^{N_2})^T \right)^T$ for every $t \geq 0$. Equation (1) can then be reformulated as

$$(5) \quad d\mathbf{Z}_t = -\lambda \mathbf{F}(\mathbf{Z}_t) dt + \sigma \mathbf{M}(\mathbf{Z}_t) d\mathbf{B}_t$$

with $(\mathbf{B}_t)_{t \geq 0}$ being a standard Brownian motion in $\mathbb{R}^{N_1 d_1 + N_2 d_2}$ and definitions

$$\begin{aligned} \mathbf{F}(\mathbf{Z}_t) &:= (F^{1,X}(\mathbf{Z}_t)^T, \dots, F^{N_1,X}(\mathbf{Z}_t)^T, F^{1,Y}(\mathbf{Z}_t)^T, \dots, F^{N_2,Y}(\mathbf{Z}_t)^T)^T \\ &\quad \text{with } F^{i,X}(\mathbf{Z}_t) = \left(X_t^i - x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}) \right) \text{ and } F^{i,Y}(\mathbf{Z}_t) = \left(Y_t^i - y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}) \right), \\ \mathbf{M}(\mathbf{Z}_t) &:= \text{diag} \left(M^{1,X}(\mathbf{Z}_t), \dots, M^{N_1,X}(\mathbf{Z}_t), M^{1,Y}(\mathbf{Z}_t), \dots, M^{N_2,Y}(\mathbf{Z}_t) \right) \\ &\quad \text{with } M^{i,X}(\mathbf{Z}_t) = D \left(X_t^i - x_\alpha^Y(\widehat{\rho}_{X,t}^{N_1}) \right) \text{ and } M^{i,Y}(\mathbf{Z}_t) = D \left(Y_t^i - y_\beta^X(\widehat{\rho}_{Y,t}^{N_2}) \right). \end{aligned}$$

The diag operator in the definition of \mathbf{M} maps the input matrices onto a block-diagonal matrix with them as its diagonal. $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ are $(N_1 d_1 + N_2 d_2) \times (N_1 d_1 + N_2 d_2)$ -dimensional diagonal matrices, whose first $N_1 d_1$ entries are λ_1 and σ_1 , and the remaining $N_2 d_2$ entries are λ_2 and σ_2 , respectively.

Having fixed the notation, we have the following well-posedness result for the SDE system (5) (respectively, (1)), which is proven towards the end of this section.

THEOREM 3. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^{d_1+d_2})$ be locally Lipschitz continuous. Then, for $N_1, N_2 \in \mathbb{N}$ fixed, the system of SDEs (1) admits a unique strong solution $(\mathbf{Z}_t)_{t \geq 0}$ for any initial condition \mathbf{Z}_0 satisfying $\mathbb{E} \|\mathbf{Z}_0\|_2^2 < \infty$.*

In order to employ the standard result [10, Chapter 5, Theorem 3.1] about the existence and uniqueness of solutions to SDEs, we need to verify that the coefficients of the SDE are locally Lipschitz continuous and of at most linear growth. This is inherited from the assumed local Lipschitz continuity of \mathcal{E} , as we make explicit in the subsequent lemma.

LEMMA 4. *Let $N_1, N_2 \in \mathbb{N}$, $\alpha, \beta > 0$, and $R > 0$ be arbitrary. Let $\mathbf{z}, \widehat{\mathbf{z}} \in \mathbb{R}^{N_1 d_1 + N_2 d_2}$ be of the form $\mathbf{z} = (\mathbf{x}^T, \mathbf{y}^T)^T = ((x^1)^T, \dots, (x^{N_1})^T, (y^1)^T, \dots, (y^{N_2})^T)^T$ and analogously for $\widehat{\mathbf{z}}$. Then, for any $\mathbf{z}, \widehat{\mathbf{z}}$ with $\|\mathbf{z}\|_2 \leq R$ and $\|\widehat{\mathbf{z}}\|_2 \leq R$, it hold for any i the bounds*

$$\|F^{i,X}(\mathbf{z})\|_2 \leq \|x^i\|_2 + \|\mathbf{x}\|_2 \quad \text{and} \quad \|F^{i,Y}(\mathbf{z})\|_2 \leq \|y^i\|_2 + \|\mathbf{y}\|_2$$

and, abbreviating $c_R(\gamma) := 4\gamma e^{2\gamma \Delta_R \mathcal{E}} \|\|\nabla_{\mathbf{z}} \mathcal{E}\|_2\|_{L^\infty(B_R)}$ with $\Delta_R \mathcal{E} := \sup_{\mathbf{z} \in B_R} \mathcal{E}(\mathbf{z}) - \inf_{\mathbf{z} \in B_R} \mathcal{E}(\mathbf{z})$,

$$\begin{aligned} \|F^{i,X}(\mathbf{z}) - F^{i,X}(\widehat{\mathbf{z}})\|_2 &\leq \|x^i - \widehat{x}^i\|_2 \\ &\quad + \left(1 + \frac{c_R(\alpha)}{N_1} \sqrt{N_1 \|\widehat{x}^i\|_2^2 + \|\widehat{\mathbf{x}}\|_2^2} \right) (\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2), \\ \|F^{i,Y}(\mathbf{z}) - F^{i,Y}(\widehat{\mathbf{z}})\|_2 &\leq \|y^i - \widehat{y}^i\|_2 \\ &\quad + \left(1 + \frac{c_R(\beta)}{N_2} \sqrt{N_2 \|\widehat{y}^i\|_2^2 + \|\widehat{\mathbf{y}}\|_2^2} \right) (\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2). \end{aligned}$$

Proof. To derive the first bound, we note that

$$\|F^{i,X}(\mathbf{z})\|_2 = \left\| x^i - \sum_{j=1}^{N_1} x^j \frac{\omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)}{\sum_{j=1}^{N_1} \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)} \right\|_2 \leq \|x^i\|_2 + \|\mathbf{x}\|_2.$$

Analogously, the bound for $\|F^{i,Y}(\mathbf{z})\|_2$ is obtained. For the other estimates, we first notice that

$$\begin{aligned} F^{i,X}(\mathbf{z}) - F^{i,X}(\widehat{\mathbf{z}}) &= \frac{\sum_{j=1}^{N_1} (x^i - x^j) \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)}{\sum_{j=1}^{N_1} \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)} \\ &\quad - \frac{\sum_{j=1}^{N_1} (\widehat{x}^i - \widehat{x}^j) \omega_\alpha(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k)}{\sum_{j=1}^{N_1} \omega_\alpha(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k)} \\ &= I_1 + I_2 + I_3 \end{aligned}$$

with I_1, I_2 , and I_3 being defined as in what follows. First, for I_1 we have

$$\begin{aligned} \|I_1\|_2 &:= \left\| \frac{\sum_{j=1}^{N_1} ((x^i - x^j) - (\widehat{x}^i - \widehat{x}^j)) \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)}{\sum_{j=1}^{N_1} \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)} \right\|_2 \\ &\leq \|x^i - \widehat{x}^i\|_2 + \|\mathbf{x} - \widehat{\mathbf{x}}\|_2. \end{aligned}$$

For I_2 and I_3 , on the other hand, let us first notice that it holds

$$\begin{aligned} &\left| \omega_\alpha\left(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k\right) - \omega_\alpha\left(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k\right) \right| \\ &\leq \alpha e^{-\alpha \inf_{\mathbf{z} \in B_R} \mathcal{E}(\mathbf{x}, \mathbf{y})} \|\nabla_{\mathbf{z}} \mathcal{E}\|_{L^\infty(B_R)} \left(\|x^j - \widehat{x}^j\|_2 + \frac{1}{N_1} \sum_{k=1}^{N_1} \|y^k - \widehat{y}^k\|_2 \right) \end{aligned}$$

and

$$\frac{1}{\sum_{j=1}^{N_1} \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)} \leq \frac{1}{N_1 \inf_{\mathbf{z} \in B_R} \exp(-\alpha \mathcal{E}(\mathbf{x}, \mathbf{y}))} \leq \frac{1}{N_1 e^{-\alpha \sup_{\mathbf{z} \in B_R} \mathcal{E}(\mathbf{x}, \mathbf{y})}}.$$

With this, we immediately obtain for the norm of I_2 the upper bound

$$\begin{aligned} \|I_2\|_2 &:= \left\| \frac{\sum_{j=1}^{N_1} (\widehat{x}^i - \widehat{x}^j) \left(\omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k) - \omega_\alpha(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k) \right)}{\sum_{j=1}^{N_1} \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)} \right\|_2 \\ &\leq \frac{2\alpha e^{\alpha \Delta_R \mathcal{E}} \|\nabla_{\mathbf{z}} \mathcal{E}\|_{L^\infty(B_R)}}{N_1} \sqrt{N_1 \|\widehat{x}^i\|_2^2 + \|\widehat{\mathbf{x}}\|_2^2} (\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2), \end{aligned}$$

where we abbreviate $\Delta_R \mathcal{E} := \sup_{\mathbf{z} \in B_R} \mathcal{E}(\mathbf{x}, \mathbf{y}) - \inf_{\mathbf{z} \in B_R} \mathcal{E}(\mathbf{x}, \mathbf{y})$. Similarly, for I_3 we have

$$\begin{aligned} \|I_3\|_2 &:= \left\| \sum_{j=1}^{N_1} (\widehat{x}^i - \widehat{x}^j) \omega_\alpha\left(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k\right) \right. \\ &\quad \times \left. \frac{\left(\sum_{j=1}^{N_1} \omega_\alpha(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k) - \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k) \right)}{\sum_{j=1}^{N_1} \omega_\alpha(\widehat{x}^j, \frac{1}{N_1} \sum_{k=1}^{N_1} \widehat{y}^k) \sum_{j=1}^{N_1} \omega_\alpha(x^j, \frac{1}{N_1} \sum_{k=1}^{N_1} y^k)} \right\|_2 \\ &\leq \frac{2\alpha e^{2\alpha \Delta_R \mathcal{E}} \|\nabla_{\mathbf{z}} \mathcal{E}\|_{L^\infty(B_R)}}{N_1} \sqrt{N_1 \|\widehat{x}^i\|_2^2 + \|\widehat{\mathbf{x}}\|_2^2} (\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|_2). \end{aligned}$$

Combining these bounds yields the result. Analogously, $\|F^{i,Y}(\mathbf{z}) - F^{i,Y}(\widehat{\mathbf{z}})\|_2$ can be bounded. \square

Proof of Theorem 3. The statement follows by invoking the standard result [10, Chapter 5, Theorem 3.1] (see Theorem A.1 in the appendix) on the existence and pathwise uniqueness of a strong solution. The fact that condition (i) of Theorem A.1 about the local Lipschitz continuity and linear growth of $\mathbf{F}(\mathbf{Z}_t)$ and $\mathbf{M}(\mathbf{Z}_t)$ holds, follows immediately from Lemma 4. To ensure condition (ii) of Theorem A.1, we make use of [10, Chapter 5, Theorem 3.2] (see Theorem A.2 in the appendix) and verify that there exists a constant $b_{N_1, N_2} > 0$ such that $-2\lambda\mathbf{Z}_t \cdot \mathbf{F}(\mathbf{Z}_t) + \text{tr}(\sigma\mathbf{M}(\mathbf{Z}_t)\mathbf{M}(\mathbf{Z}_t)^T\sigma^T) \leq b_{N_1, N_2}(1 + \|\mathbf{Z}_t\|_2^2)$. Indeed, since

$$\begin{aligned} -\lambda\mathbf{Z}_t \cdot \mathbf{F}(\mathbf{Z}_t) &\leq \lambda_1 \sum_{i=1}^{N_1} \|X_t^i\|_2 \|F^{i,X}(\mathbf{Z}_t)\|_2 + \lambda_2 \sum_{i=1}^{N_2} \|Y_t^i\|_2 \|F^{i,Y}(\mathbf{Z}_t)\|_2 \\ &\leq \left(\lambda_1(1 + \sqrt{N_1}) + \lambda_2(1 + \sqrt{N_2}) \right) \|\mathbf{Z}_t\|_2^2 \end{aligned}$$

and

$$\begin{aligned} \text{tr}(\sigma\mathbf{M}(\mathbf{Z}_t)\mathbf{M}(\mathbf{Z}_t)^T\sigma^T) &= \sigma_1^2 \sum_{i=1}^{N_1} \|F^{i,X}(\mathbf{Z}_t)\|_2^2 + \sigma_2^2 \sum_{i=1}^{N_2} \|F^{i,Y}(\mathbf{Z}_t)\|_2^2 \\ &\leq 2(\sigma_1^2(1 + N_1) + \sigma_2^2(1 + N_2)) \|\mathbf{Z}_t\|_2^2, \end{aligned}$$

the former holds with b_{N_1, N_2} defined as the sum of the two former upper bounds. \square

2.2. Well-posedness of the mean-field dynamics. In what follows, let us furthermore ensure the well-posedness of the mean-field dynamics (3) and (4), which is the main object of our studies in section 3. We prove existence and uniqueness of a solution for objective functions \mathcal{E} that satisfies the following conditions.

DEFINITION 5 (assumptions). *We consider functions $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^{d_1+d_2})$, which W1 are bounded in the sense that there exist $\underline{\mathcal{E}} \in \mathcal{C}^1(\mathbb{R}^{d_2})$ and $\bar{\mathcal{E}} \in \mathcal{C}^1(\mathbb{R}^{d_1})$ such that*

$$\underline{\mathcal{E}}(y) \leq \mathcal{E}(x, y) \leq \bar{\mathcal{E}}(x) \quad \text{for all } (x, y) \in \mathbb{R}^{d_1+d_2}.$$

W2 are locally Lipschitz continuous in the sense that there exists a constant $C_1 > 0$ such that for all $(x, y), (x', y') \in \mathbb{R}^{d_1+d_2}$ it holds that

$$|\mathcal{E}(x, y) - \mathcal{E}(x', y')| \leq C_1(1 + \|x\|_2 + \|x'\|_2 + \|y\|_2 + \|y'\|_2)(\|x - x'\|_2 + \|y - y'\|_2).$$

W3 have at most quadratic growth in the sense that there exists a constant $C_2 > 0$ obeying

$$\begin{aligned} \mathcal{E}(x, y) - \underline{\mathcal{E}}(y + sy') &\leq C_2(1 + \|x\|_2^2 + \|y\|_2^2 + \|y'\|_2^2) \quad \text{for all } (x, y), (x', y') \in \mathbb{R}^{d_1+d_2}, \quad s \in [0, 1], \\ \bar{\mathcal{E}}(x + sx') - \mathcal{E}(x, y) &\leq C_2(1 + \|x\|_2^2 + \|x'\|_2^2 + \|y\|_2^2) \quad \text{for all } (x, y), (x', y') \in \mathbb{R}^{d_1+d_2}, \quad s \in [0, 1]. \end{aligned}$$

For such objectives, we have the following well-posedness result for the macroscopic SDE (3) and its associated Fokker–Planck equation (4).

THEOREM 6. *Let \mathcal{E} satisfy assumptions W1–W3. Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$. Then there exists a unique nonlinear process $(\bar{X}, \bar{Y}) \in \mathcal{C}([0, T], \mathbb{R}^{d_1+d_2})$ satisfying (3). The associated law $\rho = \text{Law}(\bar{X}, \bar{Y})$ has regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ and is a weak solution to the Fokker–Planck equation (4).*

Before giving the proof of Theorem 6, let us first provide some auxiliary results.

LEMMA 7. Let $\varrho, \widehat{\varrho} \in \mathcal{P}_2(\mathbb{R}^{d_1+d_2})$ with $\iint \|x\|_2^2 + \|y\|_2^2 d\varrho(x, y) \leq K$ and $\iint \|\widehat{x}\|_2^2 + \|\widehat{y}\|_2^2 d\widehat{\varrho}(\widehat{x}, \widehat{y}) \leq K$. Then, under assumptions W1 and W3 on \mathcal{E} , it holds for any $s \in [0, 1]$ that

$$(6a) \quad \frac{\exp(-\alpha \mathcal{E}(\int y d\varrho_Y(y) + s(\int \widehat{y} d\widehat{\varrho}_Y(\widehat{y}) - \int y d\varrho_Y(y))))}{\int \omega_\alpha(x, \int y d\varrho_Y(y)) d\varrho_X(x)} \leq \exp(\alpha C_2(1 + 2K)) := C_K^\alpha$$

and

$$(6b) \quad \frac{\exp(\beta \bar{\mathcal{E}}(\int x d\varrho_X(x) + s(\int \widehat{x} d\widehat{\varrho}_X(\widehat{x}) - \int x d\varrho_X(x))))}{\int \omega_{-\beta}(\int x d\varrho_X(x), y) d\varrho_X(x)} \leq \exp(\beta C_2(1 + 2K)) := C_K^\beta.$$

Proof. By exploiting assumption W3 and utilizing Jensen’s inequality, we obtain

$$\begin{aligned} & \frac{\exp(-\alpha \mathcal{E}(\int y d\varrho_Y(y) + s(\int \widehat{y} d\widehat{\varrho}_Y(\widehat{y}) - \int y d\varrho_Y(y))))}{\int \omega_\alpha(x, \int y d\varrho_Y(y)) d\varrho_X(x)} \\ & \leq \frac{1}{\exp(-\alpha C_2(\int 1 + \|x\|_2^2 + \|\int y d\varrho_Y(y)\|_2^2 + \|\int \widehat{y} d\widehat{\varrho}_Y(\widehat{y})\|_2^2 d\varrho_X(x)))} \\ & \leq \exp(\alpha C_2(1 + 2K)), \end{aligned}$$

where in the last inequality we integrated the moment bounds on ϱ and $\widehat{\varrho}$. A similar estimate gives (6b) by exploiting assumption W3. \square

LEMMA 8. Let \mathcal{E} satisfy assumption W1 and the assumptions of Theorem 6. Let $\varrho, \widehat{\varrho} \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$ with $\iint \|x\|_2^4 + \|y\|_2^4 d\varrho(x, y) \leq K$ and $\iint \|\widehat{x}\|_2^4 + \|\widehat{y}\|_2^4 d\widehat{\varrho}(\widehat{x}, \widehat{y}) \leq K$. Then the stability estimate

$$(7) \quad \|x_\alpha^Y(\varrho_X) - x_\alpha^Y(\widehat{\varrho}_X)\|_2 + \|y_\beta^X(\varrho_Y) - y_\beta^X(\widehat{\varrho}_Y)\|_2 \leq c_0 W_2(\varrho, \widehat{\varrho})$$

holds with c_0 depending only on α, β, C_1, C_2 , and K .

Proof. To keep the notation concise, we write $E_{\varrho_Y} := \int y d\varrho_Y(y)$ and $E_{\widehat{\varrho}_Y} := \int \widehat{y} d\widehat{\varrho}_Y(\widehat{y})$ in what follows. According to the definition of the consensus point in (2a), we have

$$x_\alpha^Y(\varrho_X) - x_\alpha^Y(\widehat{\varrho}_X) = \iint \underbrace{\frac{x\omega_\alpha(x, E_{\varrho_Y})}{\int \omega_\alpha(x, E_{\varrho_Y}) d\varrho_X(x)} - \frac{\widehat{x}\omega_\alpha(\widehat{x}, E_{\widehat{\varrho}_Y})}{\int \omega_\alpha(\widehat{x}, E_{\widehat{\varrho}_Y}) d\widehat{\varrho}_X(\widehat{x})}}_{=: h(x) - h(\widehat{x})} d\pi(x, y, \widehat{x}, \widehat{y}),$$

where $\pi \in \Pi(\varrho, \widehat{\varrho})$ is any coupling of ϱ and $\widehat{\varrho}$. By adding and subtracting mixed terms, we obtain the decomposition

$$\begin{aligned} h(x) - h(\widehat{x}) &= \frac{(x - \widehat{x})\omega_\alpha(x, E_{\varrho_Y})}{\int \omega_\alpha(x, E_{\varrho_Y}) d\varrho_X(x)} + \frac{\widehat{x}(\omega_\alpha(x, E_{\varrho_Y}) - \omega_\alpha(\widehat{x}, E_{\widehat{\varrho}_Y}))}{\int \omega_\alpha(x, E_{\varrho_Y}) d\varrho_X(x)} \\ & \quad + \frac{\int \omega_\alpha(\widehat{x}, E_{\widehat{\varrho}_Y}) - \omega_\alpha(x, E_{\varrho_Y}) d\pi(x, y, \widehat{x}, \widehat{y})}{(\int \omega_\alpha(x, E_{\varrho_Y}) d\varrho_X(x)) (\int \omega_\alpha(\widehat{x}, E_{\widehat{\varrho}_Y}) d\widehat{\varrho}_X(\widehat{x}))} \widehat{x}\omega_\alpha(\widehat{x}, E_{\widehat{\varrho}_Y}) \\ & =: I_1 + I_2 + I_3, \end{aligned}$$

where I_1, I_2 , and I_3 correspond to the three summands. For I_1 , by using assumption W3 and Lemma 7 with $s=0$, we obtain

$$\|I_1\|_2 \leq \frac{\omega_\alpha(x, \mathbf{E}\varrho_Y)}{\int \omega_\alpha(x, \mathbf{E}\varrho_Y) d\varrho_X(x)} \|x - \hat{x}\|_2 \leq \frac{e^{-\alpha\xi(\mathbf{E}\varrho_Y)}}{\int \omega_\alpha(x, \mathbf{E}\varrho_Y) d\varrho_X(x)} \|x - \hat{x}\|_2 \leq C_K^\alpha \|x - \hat{x}\|_2.$$

For I_2 and I_3 , on the other hand, let us first notice that for some $s, s' \in [0, 1]$ it holds that

$$\begin{aligned} |\omega_\alpha(x, \mathbf{E}\varrho_Y) - \omega_\alpha(\hat{x}, \mathbf{E}\hat{\varrho}_Y)| &\leq |\omega_\alpha(x, \mathbf{E}\varrho_Y) - \omega_\alpha(\hat{x}, \mathbf{E}\varrho_Y)| + |\omega_\alpha(\hat{x}, \mathbf{E}\varrho_Y) - \omega_\alpha(\hat{x}, \mathbf{E}\hat{\varrho}_Y)| \\ &= |\partial_x \omega_\alpha(x + s(\hat{x} - x), \mathbf{E}\varrho_Y)| \|x - \hat{x}\|_2 \\ &\quad + |\partial_y \omega_\alpha(\hat{x}, \mathbf{E}\varrho_Y + s'(\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y))| \|\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y\|_2 \\ &\leq \alpha C_1 e^{-\alpha\xi(\mathbf{E}\varrho_Y)} 2 \left(1 + \|x\|_2 + \|\hat{x}\|_2 + \|\mathbf{E}\varrho_Y\|_2\right) \|x - \hat{x}\|_2 \\ &\quad + \alpha C_1 e^{-\alpha\xi(\mathbf{E}\varrho_Y + s'(\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y))} 2 \\ &\quad \times \left(1 + \|\mathbf{E}\hat{\varrho}_Y\|_2 + \|\hat{x}\|_2 + \|\mathbf{E}\varrho_Y\|_2\right) \|\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y\|_2 \end{aligned}$$

due to assumptions W2 and W3. With this, we immediately obtain the upper bounds

$$\begin{aligned} \|I_2\|_2 &\leq 2\alpha C_1 C_K^\alpha \|\hat{x}\|_2 \left(1 + \|x\|_2 + \|\mathbf{E}\hat{\varrho}_Y\|_2 + 2\|\hat{x}\|_2 + 2\|\mathbf{E}\varrho_Y\|_2\right) (\|x - \hat{x}\|_2 \\ &\quad + \|\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y\|_2), \\ \|I_3\|_2 &\leq 2\alpha C_1 (C_K^\alpha)^2 \|\hat{x}\|_2 \cdot \iint \left(1 + \|x\|_2 + \|\mathbf{E}\hat{\varrho}_Y\|_2 + 2\|\hat{x}\|_2 + 2\|\mathbf{E}\varrho_Y\|_2\right) \\ &\quad \times (\|x - \hat{x}\|_2 + \|\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y\|_2) d\pi(x, y, \hat{x}, \hat{y}). \end{aligned}$$

Collecting the latter three estimates for $\|I_1\|_2, \|I_2\|_2$, and $\|I_3\|_2$ eventually gives after an application of Jensen's and Cauchy–Schwarz's inequalities

$$\begin{aligned} \|x_\alpha^Y(\varrho_X) - x_\alpha^Y(\hat{\varrho}_X)\|_2 &\leq C(\alpha, C_1, C_K^\alpha, K) \sqrt{\iint \|x - \hat{x}\|_2^2 d\pi(x, y, \hat{x}, \hat{y}) + \|\mathbf{E}\hat{\varrho}_Y - \mathbf{E}\varrho_Y\|_2^2} \\ &\leq C(\alpha, C_1, C_K^\alpha, K) \sqrt{\iint \|x - \hat{x}\|_2^2 + \|y - \hat{y}\|_2^2 d\pi(x, y, \hat{x}, \hat{y})}, \end{aligned}$$

where the last step is due to Jensen's inequality. $\|y_\beta^X(\varrho_Y) - y_\beta^X(\hat{\varrho}_Y)\|_2$ can be bounded analogously. Eventually, optimizing over all couplings $\pi \in \Pi(\varrho, \hat{\varrho})$ gives the claim. \square

Proof of Theorem 6. The proof is based on the Leray–Schauder fixed point theorem and follows in the spirit of [5, Theorems 3.1, 3.2].

Step 1. For given functions $u^X \in \mathcal{C}([0, T], \mathbb{R}^{d_1})$, $u^Y \in \mathcal{C}([0, T], \mathbb{R}^{d_2})$ and measure $\rho_0 \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$, according to standard SDE theory [1, Chapter 6], we can uniquely solve the SDE system

$$(8a) \quad d\tilde{X}_t = -\lambda_1(\tilde{X}_t - u_t^X) dt + \sigma_1 D(\tilde{X}_t - u_t^X) dB_t^X,$$

$$(8b) \quad d\tilde{Y}_t = -\lambda_2(\tilde{Y}_t - u_t^Y) dt + \sigma_2 D(\tilde{Y}_t - u_t^Y) dB_t^Y$$

with $(\tilde{X}_0, \tilde{Y}_0) \sim \rho_0$ as a consequence of the coefficients being locally Lipschitz continuous and having at most linear growth. This induces $\tilde{\rho}_t = \text{Law}((\tilde{X}_t, \tilde{Y}_t))$. Moreover, the regularity of the initial distribution $\rho_0 \in \mathcal{P}_4(\mathbb{R}^{d_1+d_2})$ allows for a fourth-order moment estimate of the form $\mathbb{E}[\|\tilde{X}_t\|_2^4 + \|\tilde{Y}_t\|_2^4] \leq (1 + 2\mathbb{E}[\|\tilde{X}_0\|_2^4 + \|\tilde{Y}_0\|_2^4])e^{ct}$; see, e.g., [1, Chapter 7]. So, in particular, $\tilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$, i.e., $\sup_{t \in [0, T]} \iint \|x\|_2^4 + \|y\|_2^4 d\tilde{\rho}_t(x, y) \leq K$ for some $K > 0$.

Step 2. Let us now define the test function space

$$(9) \quad \mathcal{C}_*^2(\mathbb{R}^{d_1+d_2}) := \left\{ \phi \in \mathcal{C}^2(\mathbb{R}^{d_1+d_2}) : \|\nabla\phi(x, y)\|_2 \leq C_\phi(1 + \|x\|_2 + \|y\|_2) \text{ for some } C_\phi > 0 \right. \\ \left. \text{and } \max \left\{ \max_{k=1, \dots, d_1} \|\partial_{x_k x_k}^2 \phi\|_\infty, \max_{k=1, \dots, d_2} \|\partial_{y_k y_k}^2 \phi\|_\infty \right\} < \infty \right\}.$$

For any $\phi \in \mathcal{C}_*^2(\mathbb{R}^d)$, by Itô's formula, we can derive

$$d\phi(\tilde{X}_t, \tilde{Y}_t) = \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot \left(-\lambda_1(\tilde{X}_t - u_t^X) dt + \sigma_1 D(\tilde{X}_t - u_t^X) dB_t^X \right) \\ + \nabla_y \phi(\tilde{X}_t, \tilde{Y}_t) \cdot \left(-\lambda_2(\tilde{Y}_t - u_t^Y) dt + \sigma_2 D(\tilde{Y}_t - u_t^Y) dB_t^Y \right) \\ + \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \partial_{x_k x_k}^2 \phi(\tilde{X}_t, \tilde{Y}_t) (\tilde{X}_t - u_t^X)_k^2 dt \\ + \frac{\sigma_2^2}{2} \sum_{k=1}^{d_2} \partial_{y_k y_k}^2 \phi(\tilde{X}_t, \tilde{Y}_t) (\tilde{Y}_t - u_t^Y)_k^2 dt.$$

After taking the expectation, applying Fubini's theorem, and observing that the stochastic integrals $\mathbb{E} \int_0^t \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot D(\tilde{X}_t - u_t^X) dB_t^X$ and $\mathbb{E} \int_0^t \nabla_y \phi(\tilde{X}_t, \tilde{Y}_t) \cdot D(\tilde{Y}_t - u_t^Y) dB_t^Y$ vanish as a consequence of [37, Theorem 3.2.1(iii)] due to the established regularity $\tilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ and $\phi \in \mathcal{C}_*^2(\mathbb{R}^{d_1+d_2})$, we obtain

$$\frac{d}{dt} \mathbb{E} \phi(\tilde{X}_t, \tilde{Y}_t) = -\lambda_1 \mathbb{E} \nabla_x \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (\tilde{X}_t - u_t^X) - \lambda_2 \mathbb{E} \nabla_y \phi(\tilde{X}_t, \tilde{Y}_t) \cdot (\tilde{Y}_t - u_t^Y) \\ + \frac{\sigma_1^2}{2} \mathbb{E} \sum_{k=1}^{d_1} \partial_{x_k x_k}^2 \phi(\tilde{X}_t, \tilde{Y}_t) (\tilde{X}_t - u_t^X)_k^2 \\ + \frac{\sigma_2^2}{2} \mathbb{E} \sum_{k=1}^{d_2} \partial_{y_k y_k}^2 \phi(\tilde{X}_t, \tilde{Y}_t) (\tilde{Y}_t - u_t^Y)_k^2$$

as a consequence of the fundamental theorem of calculus. This shows that the measure $\tilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ satisfies the Fokker–Planck equation

$$(10) \quad \begin{aligned} \frac{d}{dt} \int \phi(x, y) d\tilde{\rho}_t(x, y) &= - \int \lambda_1 \nabla_x \phi(x, y) \cdot (x - u_t^X) + \lambda_2 \nabla_y \phi(x, y) \cdot (y - u_t^Y) d\tilde{\rho}_t(x, y) \\ &\quad + \int \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \partial_{x_k x_k}^2 \phi(x, y) (x - u_t^X)_k^2 \\ &\quad + \frac{\sigma_2^2}{2} \sum_{k=1}^{d_2} \partial_{y_k y_k}^2 \phi(x, y) (y - u_t^Y)_k^2 d\tilde{\rho}_t(x, y). \end{aligned}$$

Step 3. Setting $\mathcal{T}u := ((x_\alpha^Y(\tilde{\rho}_X))^T, (y_\beta^X(\tilde{\rho}_Y))^T)^T \in \mathcal{C}([0, T], \mathbb{R}^{d_1+d_2})$ provides the self-mapping property of the map $\mathcal{T} : \mathcal{C}([0, T], \mathbb{R}^{d_1+d_2}) \rightarrow \mathcal{C}([0, T], \mathbb{R}^{d_1+d_2})$, $u \mapsto \mathcal{T}u = ((x_\alpha^Y(\tilde{\rho}_X))^T, (y_\beta^X(\tilde{\rho}_Y))^T)^T$. By means of Lemma 8, we have

$$\|x_\alpha^Y(\tilde{\rho}_{X,t}) - x_\alpha^Y(\tilde{\rho}_{X,s})\|_2 + \|y_\beta^X(\tilde{\rho}_{Y,t}) - y_\beta^X(\tilde{\rho}_{Y,s})\|_2 \leq c_0 W_2(\tilde{\rho}_t, \tilde{\rho}_s) \lesssim c_0 |t - s|^{1/2},$$

which shows the Hölder-1/2 continuity of \mathcal{T} due to the compact embedding $\mathcal{C}^{0,1/2}([0, T], \mathbb{R}^{d_1+d_2}) \hookrightarrow \mathcal{C}([0, T], \mathbb{R}^{d_1+d_2})$. In the last inequality, we note that due to Itô’s isometry it holds that

$$\begin{aligned} W_2^2(\tilde{\rho}_t, \tilde{\rho}_s) &\leq \mathbb{E} \left[\|\tilde{X}_t - \tilde{X}_s\|_2^2 + \|\tilde{Y}_t - \tilde{Y}_s\|_2^2 \right] \\ &\leq 4 \left((\lambda_1^2 + \lambda_2^2)T + (\sigma_1^2 + \sigma_2^2) \right) \left(K + \|u^X\|_{L^\infty}^2 + \|u^Y\|_{L^\infty}^2 \right) |t - s|. \end{aligned}$$

Step 4. Now, for $u \in \mathcal{C}([0, T], \mathbb{R}^{d_1+d_2})$ satisfying $u = \vartheta \mathcal{T}u$ with $\vartheta \in [0, 1]$, there exists $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{d_1+d_2}))$ satisfying (10) such that $u = \vartheta ((x_\alpha^Y(\rho_X))^T, (y_\beta^X(\rho_Y))^T)^T$. As a consequence of Lemma 7, with $s = 0$ we can show that

$$\begin{aligned} \|u_t\|_2^2 &\leq \vartheta^2 \left((C_K^\alpha)^2 \int \|x\|_2^2 d\rho_{X,t}(x) + (C_K^\beta)^2 \int \|y\|_2^2 d\rho_{Y,t}(y) \right) \\ &\leq \vartheta^2 \left((C_K^\alpha)^2 + (C_K^\beta)^2 \right) \sqrt{K}, \end{aligned}$$

where we can use $K = (1 + 2\mathbb{E}[\|\tilde{X}_0\|_2^4 + \|\tilde{Y}_0\|_2^4])e^{cT}$ of Step 1. This allows for a uniform estimate of $\|u\|_{L^\infty} < q$ for $q > 0$. An application of the Leray–Schauder fixed point theorem concludes the proof of existence by providing a solution to (3).

Step 5. For uniqueness, suppose we have two fixed points u^1 and u^2 (as specified in the previous step) together with corresponding processes $((\tilde{X}^1)^T, (\tilde{Y}^1)^T)^T$ and $((\tilde{X}^2)^T, (\tilde{Y}^2)^T)^T$ satisfying (8). Then, taking their difference while keeping the initial conditions and respective Brownian motion paths, we obtain after an application of Itô’s isometry and the employment of Lemma 8 the bound

$$\begin{aligned} \mathbb{E} \left[\|\tilde{X}_t^1 - \tilde{X}_t^2\|_2^2 + \|\tilde{Y}_t^1 - \tilde{Y}_t^2\|_2^2 \right] &\leq c \mathbb{E} \int_0^t \left(\|\tilde{X}_\tau^1 - \tilde{X}_\tau^2\|_2^2 + \|\tilde{Y}_\tau^1 - \tilde{Y}_\tau^2\|_2^2 + \|x_\alpha^Y(\tilde{\rho}_{X,\tau}^1) \right. \\ &\quad \left. - x_\alpha^Y(\tilde{\rho}_{X,\tau}^2)\|_2^2 + \|y_\beta^X(\tilde{\rho}_{Y,\tau}^1) - y_\beta^X(\tilde{\rho}_{Y,\tau}^2)\|_2^2 \right) d\tau \\ &\lesssim c \mathbb{E} \int_0^t \left(\|\tilde{X}_\tau^1 - \tilde{X}_\tau^2\|_2^2 + \|\tilde{Y}_\tau^1 - \tilde{Y}_\tau^2\|_2^2 \right) d\tau \end{aligned}$$

with $c = 4 \left((\lambda_1^2 + \lambda_2^2)T + (\sigma_1^2 + \sigma_2^2) \right)$. Grönwall’s inequality eventually shows uniqueness since $\mathbb{E}[\|\tilde{X}_t^1 - \tilde{X}_t^2\|_2^2 + \|\tilde{Y}_t^1 - \tilde{Y}_t^2\|_2^2] = 0$ for all $t \in [0, T]$. \square

3. Convergence to saddle points. Inspired by the theories of mean-field limits for consensus-based optimization methods (see [11, 14, 20], for instance), the convergence of particles systems (1) to the mean-field dynamics (3) follows in a similar way, and thus the associated argument is omitted. In this section, we present the main theoretical result of our paper concerned with the convergence of the macroscopic dynamics (3) towards saddle points of objective functions \mathcal{E} that satisfy the following conditions.

DEFINITION 9 (assumptions). *Throughout this section, we are interested in objective functions $\mathcal{E} \in C^2(\mathbb{R}^{d_1+d_2})$, for which*

A1 *there exist two functions $\underline{\mathcal{E}} \in C^1(\mathbb{R}^{d_2})$ and $\bar{\mathcal{E}} \in C^1(\mathbb{R}^{d_1})$ such that*

$$\underline{\mathcal{E}}(y) \leq \mathcal{E}(x, y) \leq \bar{\mathcal{E}}(x)$$

for all $(x, y) \in \mathbb{R}^{d_1+d_2}$. The functions $\underline{\mathcal{E}}$ and $\bar{\mathcal{E}}$ shall, for some constant $\bar{C}_{\nabla\mathcal{E}} > 0$, satisfy $\|\nabla\underline{\mathcal{E}}(y)\|_2 \leq \bar{C}_{\nabla\mathcal{E}}$ for all $y \in \mathbb{R}^{d_2}$ and $\|\nabla\bar{\mathcal{E}}(x)\|_2 \leq \bar{C}_{\nabla\mathcal{E}}$ for all $x \in \mathbb{R}^{d_1}$.

A2 *there exist constants $C_{\nabla\mathcal{E}}, C_{\nabla^2\mathcal{E}} > 0$ such that*

$$\max \left\{ \sup_{(x,y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} \|\nabla_x \mathcal{E}(x, y)\|_2, \sup_{(x,y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} \|\nabla_y \mathcal{E}(x, y)\|_2 \right\} \leq C_{\nabla\mathcal{E}},$$

$$\max \left\{ \max_{k=1, \dots, d_1} \|\partial_{x_k x_k}^2 \mathcal{E}\|_\infty, \max_{k=1, \dots, d_2} \|\partial_{y_k y_k}^2 \mathcal{E}\|_\infty, \|\rho(\nabla_x^2 \mathcal{E})\|_\infty, \|\rho(\nabla_y^2 \mathcal{E})\|_\infty \right\} \leq C_{\nabla^2\mathcal{E}},$$

where $\|\cdot\|_\infty$ denotes the L^∞ norm on $\mathcal{C}(\mathbb{R}^{d_1+d_2})$ and ρ denotes the spectral radius.

A3 *there exist constants $\epsilon_0, \eta, \nu > 0$ such that for each $(x, y) \in \mathbb{R}^{d_1+d_2}$ satisfying $\mathcal{E}^* - \mathcal{E}(x^*, y) \leq \epsilon_0$ and $\mathcal{E}(x, y^*) - \mathcal{E}^* \leq \epsilon_0$ for some saddle point (x^*, y^*) of \mathcal{E} , we have*

$$\|x - x^*\|_2 \leq \frac{1}{\eta} (|\mathcal{E}(x, y^*) - \mathcal{E}^*|)^\nu \quad \text{and} \quad \|y - y^*\|_2 \leq \frac{1}{\eta} (|\mathcal{E}(x^*, y) - \mathcal{E}^*|)^\nu.$$

Assumption A1 requires that the twice continuously differentiable objective function \mathcal{E} is bounded from below by a function $\underline{\mathcal{E}}$, which depends only on the y -coordinate, and from above by a function $\bar{\mathcal{E}}$, which depends only on the x -coordinate. Moreover, the first-order derivatives of $\underline{\mathcal{E}}$ and $\bar{\mathcal{E}}$ are assumed to be uniformly bounded.

Assumption A2 is a mere technical regularity assumption about \mathcal{E} in terms of the first and second derivatives. In particular, it requires that the gradients as well as the second-order derivatives of \mathcal{E} are uniformly bounded, which is, however, necessary only for theoretical analysis of the long time behavior of the algorithm. Analogous regularity assumptions may be found in the literature; see, e.g., [5, 6, 21]. However, as a purely zero-order derivative-free method, our CBO-SP algorithm only uses pointwise values of the objective function \mathcal{E} in practical applications.

Assumption A3, on the other hand, should be regarded as a tractability condition on the landscape of the objective function \mathcal{E} . It imposes coercivity of \mathcal{E} around saddle points, which relates the distance from (x^*, y^*) with the value of the objective function. We refer the reader to the discussion after [14, Remark 9] for related conditions in the machine learning literature.

In order to formulate in Theorem 11 below the result about the convergence of the dynamics (3) towards saddle points of the objective functions \mathcal{E} satisfying the aforementioned assumption, let us define the variances

$$(11) \quad \text{Var}^X(t) = \mathbb{E} \|\bar{X}_t - \mathbb{E}\bar{X}_t\|_2^2 \quad \text{and} \quad \text{Var}^Y(t) = \mathbb{E} \|\bar{Y}_t - \mathbb{E}\bar{Y}_t\|_2^2,$$

which act as Lyapunov functionals. In addition, we require certain well-preparedness assumptions about the initial data and parameters. For this reason, we introduce the notations

$$\begin{aligned}
 (12a) \quad & \widetilde{\mathcal{M}}^X(t) := \mathbb{E} \exp(-\alpha \mathcal{E}(\overline{X}_t, \mathbb{E}\overline{Y}_t)) \quad \text{and} \quad \widetilde{\mathcal{M}}^Y(t) := \mathbb{E} \exp(\beta \mathcal{E}(\mathbb{E}\overline{X}_t, \overline{Y}_t)), \\
 (12b) \quad & \mathcal{M}^X(t) := \widetilde{\mathcal{M}}^X(t) e^{\alpha \underline{\mathcal{E}}(\mathbb{E}\overline{Y}_t)} \quad \text{and} \quad \mathcal{M}^Y(t) := \widetilde{\mathcal{M}}^Y(t) e^{-\beta \overline{\mathcal{E}}(\mathbb{E}\overline{X}_t)}, \\
 (12c) \quad & \mathcal{M}_*^X(t) := \mathbb{E} \exp(-\alpha \mathcal{E}(\overline{X}_t, y^*)) \quad \text{and} \quad \mathcal{M}_*^Y(t) := \mathbb{E} \exp(\beta \mathcal{E}(x^*, \overline{Y}_t)),
 \end{aligned}$$

where the definitions of \mathcal{M}_*^X and \mathcal{M}_*^Y in (12c) may be different for potentially different saddle points (x^*, y^*) (since our assumptions allow for the existence of multiple such points), meaning, in particular, that there may be multiple functionals \mathcal{M}_*^X and \mathcal{M}_*^Y . This ambiguity is clarified in Remark 12.

DEFINITION 10 (well-preparedness of the initial data and parameters). *The initial datum $(\overline{X}_0, \overline{Y}_0)$ and the parameters $\alpha, \beta, \lambda_1, \lambda_2, \sigma_1$, and σ_2 of the CBO-SP method are well-prepared if*

- P1 $\mu_1 := 2(\lambda_1 - 4\sigma_1^2/\mathcal{M}^X(0)) > 0$ and $\mu_2 := 2(\lambda_2 - 4\sigma_2^2/\mathcal{M}^Y(0)) > 0$.
- P2 all saddle points (x^*, y^*) lie in $\text{supp}(\rho_0)$, where $\rho_0 := \text{Law}(\overline{X}_0, \overline{Y}_0)$ has marginals ρ_0^X and ρ_0^Y . Moreover, for any $\delta > 0$, there exists some constant $C_\delta > 0$ depending only on δ such that it holds that

$$\begin{aligned}
 \rho_0^X \left(\left\{ x : \exp(-\mathcal{E}(x, \mathbb{E}(\overline{Y}_0))) > \exp\left(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}(\overline{Y}_0))\right) - \delta \right\} \right) &\geq C_\delta, \\
 \rho_0^Y \left(\left\{ y : \exp(\mathcal{E}(\mathbb{E}(\overline{X}_0), y)) > \exp\left(\max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}(\overline{X}_0), y)\right) - \delta \right\} \right) &\geq C_\delta
 \end{aligned}$$

as well as

$$\begin{aligned}
 \rho_0^X(\{x : \exp(-\mathcal{E}(x, y^*)) > \exp(-\mathcal{E}^*) - \delta\}) &\geq C_\delta, \\
 \rho_0^Y(\{y : \exp(\mathcal{E}(x^*, y)) > \exp(\mathcal{E}^*) - \delta\}) &\geq C_\delta.
 \end{aligned}$$

- P3 it holds that

$$\begin{aligned}
 4\alpha C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(0)}{\mu_1 \mathcal{M}^X(0)} + 2\sqrt{2}\alpha\lambda_2 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)}}{\mu_2 \sqrt{\mathcal{M}^Y(0)}} &\leq \frac{1}{8} \mathcal{M}^X(0), \\
 4\beta C_{\nabla^2 \mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(0)}{\mu_2 \mathcal{M}^Y(0)} + 2\sqrt{2}\beta\lambda_1 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(0)}}{\mu_1 \sqrt{\mathcal{M}^X(0)}} &\leq \frac{1}{8} \mathcal{M}^Y(0).
 \end{aligned}$$

- P4 it holds for any fixed (x^*, y^*) that

$$\begin{aligned}
 &4\alpha\sigma_1^2 C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(0)}{\mu_1 \mathcal{M}^X(0)} + 8\alpha\lambda_1 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(0)}}{\mu_1 \sqrt{\mathcal{M}^X(0)}} + \sqrt{2}\alpha\lambda_2 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)}}{\mu_2 \sqrt{\mathcal{M}^Y(0)}} \\
 &\leq \frac{1}{8} \min \{ \mathcal{M}_*^X(0) e^{\alpha \underline{\mathcal{E}}(y^*)}, \widetilde{\mathcal{M}}^X(0) e^{-\alpha \mathcal{E}_M} \}, \\
 &4\beta C_{\nabla^2 \mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(0)}{\mu_2 \mathcal{M}^Y(0)} + 8\beta\lambda_2 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)}}{\mu_2 \sqrt{\mathcal{M}^Y(0)}} + \sqrt{2}\beta\lambda_1 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(0)}}{\mu_1 \sqrt{\mathcal{M}^X(0)}} \\
 &\leq \frac{1}{8} \min \{ \mathcal{M}_*^Y(0) e^{-\beta \overline{\mathcal{E}}(x^*)}, \widetilde{\mathcal{M}}^Y(0) e^{-\beta \mathcal{E}_M} \},
 \end{aligned}$$

where \mathcal{E}_M depends on (\tilde{x}, \tilde{y}) as well as $\underline{\mathcal{E}}$ and $\overline{\mathcal{E}}$ (see Theorem 11 below).

Condition P1 can be satisfied appropriately for big choices of λ_1 and λ_2 . Condition P2 is valid if the initial distribution ρ_0 has some mass at the saddle points (x^*, y^*) . While this may have a certain locality flavor, in the case of functions \mathcal{E} having multiple saddle points, the condition is generically satisfied at least for one of them, essentially allowing us to obtain a global result. It is actually sufficient if there is at least one saddle point satisfying condition P2, in which case the method is agnostic to saddle points not in $\text{supp}(\rho_0)$. Conditions P3–P4, on the other hand, may be ensured if the initial variances $\text{Var}^X(0)$ and $\text{Var}^Y(0)$ are sufficiently small. Well-preparedness conditions similar to P2–P4 can be found in the literature about a convergence analysis of CBO for minimization (see, e.g., [5, 6, 12, 21]), while we note that the coupling of (\bar{X}, \bar{Y}) due to the intrinsic difference between games and optimizations prompts us to use different proof techniques.

We are now ready to state the result about the convergence of the dynamics (3) towards saddle points of the objective functions \mathcal{E} . The proof details are deferred to section 4.

THEOREM 11. *Let \mathcal{E} satisfy assumptions A1 and A2, and let $(\bar{X}_t, \bar{Y}_t)_{t \geq 0}$ be a solution to the SDE (3). Then the following statements hold:*

- (1) *Under the assumption of well-preparedness of the initial datum (\bar{X}_0, \bar{Y}_0) and the parameters $\alpha, \beta, \lambda_1, \lambda_2, \sigma_1$, and σ_2 in the sense of P1–P3, Var^X and Var^Y as defined in (11) converge exponentially fast to 0 as $t \rightarrow \infty$. More precisely, it holds that*

$$(13) \quad \text{Var}^X(t) + \text{Var}^Y(t) \leq \text{Var}^X(0)e^{-\mu_1 t} + \text{Var}^Y(0)e^{-\mu_2 t}.$$

Moreover, there exists some (\tilde{x}, \tilde{y}) depending in particular on α and β such that, as $t \rightarrow \infty$,

$$(14) \quad (\mathbb{E}\bar{X}_t, \mathbb{E}\bar{Y}_t) \rightarrow (\tilde{x}, \tilde{y}) \quad \text{and} \quad (x_\alpha^Y(\rho_{X,t}), y_\beta^X(\rho_{Y,t})) \rightarrow (\tilde{x}, \tilde{y}).$$

- (2) *For any given accuracy $\varepsilon > 0$, there exist some $\alpha_0, \beta_0 > 0$ such that for all $\alpha \geq \alpha_0$ and $\beta \geq \beta_0$ the point (\tilde{x}, \tilde{y}) from (1) (which may depend on α and β) satisfies*

$$(15) \quad |\mathcal{E}(\tilde{x}, \tilde{y}) - \mathcal{E}^*| \leq \varepsilon \quad \text{as well as} \quad \mathcal{E}^* - \mathcal{E}(x^*, \tilde{y}) \leq \varepsilon \quad \text{and} \quad \mathcal{E}(\tilde{x}, y^*) - \mathcal{E}^* \leq \varepsilon$$

provided that the well-preparedness assumptions P1–P4 hold for such α and β together with the initial datum (\bar{X}_0, \bar{Y}_0) .

- (3) *If \mathcal{E} satisfies assumption A3 with respect to (\tilde{x}, \tilde{y}) from (2) with $\varepsilon \leq \varepsilon_0$, i.e., there exists some saddle point (x^*, y^*) depending on (\tilde{x}, \tilde{y}) such that $\|\tilde{x} - x^*\|_2 \leq \frac{1}{\eta} (|\mathcal{E}(\tilde{x}, y^*) - \mathcal{E}^*|)^\nu$ and $\|\tilde{y} - y^*\|_2 \leq \frac{1}{\eta} (|\mathcal{E}(x^*, \tilde{y}) - \mathcal{E}^*|)^\nu$, then we have*

$$(16) \quad \|(\tilde{x}, \tilde{y}) - (x^*, y^*)\|_2 \leq \frac{2}{\eta} \varepsilon^\nu$$

provided that the well-preparedness assumptions P1–P4 hold for sufficiently large α and β together with the initial datum (\bar{X}_0, \bar{Y}_0) .

Part (1) of Theorem 11 states that under suitable well-preparedness conditions on the initialization and the parameters, the mean-field dynamics (3) reaches consensus at *some* location (\tilde{x}, \tilde{y}) , which may depend in particular on α and β , as time evolves. In part (2) of the statement, for sufficiently large α and β as well as under certain

well-preparedness conditions, properties of *the* corresponding point (\tilde{x}, \tilde{y}) are specified, which are typical for saddle points; see (15). These properties eventually allow one to conclude part (3) of the result, that *the* (\tilde{x}, \tilde{y}) from before is arbitrarily close to *any* saddle point (x^*, y^*) which satisfies the inverse continuity property A3.

Remark 12. It is worth mentioning at this point that in order to prove (15), any saddle point (x^*, y^*) satisfying assumption P4 can be used in the definitions of \mathcal{M}_*^X and \mathcal{M}_*^Y . For the proof of (16), on the other hand, it is necessary to use in the definitions of \mathcal{M}_*^X and \mathcal{M}_*^Y a specific saddle point (x^*, y^*) that satisfies the inverse continuity property A3 with respect to (\tilde{x}, \tilde{y}) as well as assumption P4, so, in this case, the saddle point (x^*, y^*) does depend on (\tilde{x}, \tilde{y}) .

4. Proof details for section 3. In this section, we provide the proof details for the convergence result of the mean-field dynamics (3) to a saddle point of the objective function \mathcal{E} . Sections 4.1–4.3 present individual statements which are necessary in the proof of our main theorem, Theorem 11, which is then given in section 4.4.

4.1. Time-evolution of the variances Var^X and Var^Y . In order to ensure consensus formation of the mean-field dynamics (3), we show that the variances $\text{Var}^X(t) = \mathbb{E}\|\bar{X}_t - \mathbb{E}\bar{X}_t\|_2^2$ and $\text{Var}^Y(t) = \mathbb{E}\|\bar{Y}_t - \mathbb{E}\bar{Y}_t\|_2^2$ of the particle distribution decay to 0 as $t \rightarrow \infty$. For this, we need to analyze their time evolutions, as done in what follows.

LEMMA 13. *Let Var^X and Var^Y be as defined in (11), and let us recall the definitions of \mathcal{M}^X and \mathcal{M}^Y from (12b). Then, under assumption A1, it holds that*

$$(17) \quad \begin{aligned} \frac{d}{dt} \text{Var}^X(t) &\leq -2 \left(\lambda_1 - \frac{2\sigma_1^2}{\mathcal{M}^X(t)} \right) \text{Var}^X(t) \quad \text{and} \\ \frac{d}{dt} \text{Var}^Y(t) &\leq -2 \left(\lambda_2 - \frac{2\sigma_2^2}{\mathcal{M}^Y(t)} \right) \text{Var}^Y(t). \end{aligned}$$

Proof. By means of Itô’s calculus, $d\|\bar{X}_t - \mathbb{E}\bar{X}_t\|_2^2 = 2(\bar{X}_t - \mathbb{E}\bar{X}_t) \cdot d\bar{X}_t + \sigma_1^2 \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2 dt$. Since the appearing stochastic integral vanishes when taking the expectation as a consequence of the regularity established in Theorem 6 and assumption A1, we obtain

$$(18) \quad \begin{aligned} \frac{d}{dt} \text{Var}^X(t) &= -2\lambda_1 \mathbb{E}[(\bar{X}_t - \mathbb{E}\bar{X}_t) \cdot (\bar{X}_t - x_\alpha^Y(\rho_{X,t}))] + \sigma_1^2 \mathbb{E}\|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2 \\ &= -2\lambda_1 \text{Var}^X(t) + \sigma_1^2 \mathbb{E}\|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2, \end{aligned}$$

where we used that $\mathbb{E}[(\bar{X}_t - \mathbb{E}\bar{X}_t) \cdot (\mathbb{E}\bar{X}_t - x_\alpha^Y(\rho_{X,t}))] = 0$. Analogously, we derive

$$(19) \quad \frac{d}{dt} \text{Var}^Y(t) = -2\lambda_2 \text{Var}^Y(t) + \sigma_2^2 \mathbb{E}\|\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2^2.$$

In order to control the terms $\mathbb{E}\|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2$ and $\mathbb{E}\|\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2^2$ appearing in (18) and (19), let us first observe that, for any $\hat{x} \in \mathbb{R}^{d_1}$, $\hat{y} \in \mathbb{R}^{d_2}$, Jensen’s inequality gives

$$(20) \quad \|\hat{x} - x_\alpha^Y(\rho_{X,t})\|_2^2 \leq \frac{1}{\mathbb{E}\omega_\alpha(\bar{X}_t, \mathbb{E}\bar{Y}_t)} \int \|\hat{x} - x\|_2^2 \omega_\alpha(x, \mathbb{E}\bar{Y}_t) d\rho_{X,t}(x),$$

$$(21) \quad \|\hat{y} - y_\beta^X(\rho_{Y,t})\|_2^2 \leq \frac{1}{\mathbb{E}\omega_{-\beta}(\mathbb{E}\bar{X}_t, \bar{Y}_t)} \int \|\hat{y} - y\|_2^2 \omega_{-\beta}(\mathbb{E}\bar{X}_t, y) d\rho_{Y,t}(y).$$

Exploiting the boundedness of \mathcal{E} as of assumption A1, the two latter bounds in particular imply

$$(22) \quad \begin{aligned} \mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2 &\leq \frac{2}{\mathbb{E}\omega_\alpha(\bar{X}_t, \mathbb{E}\bar{Y}_t)} \\ &\quad \times \int \left(\mathbb{E} \|\bar{X}_t - \mathbb{E}\bar{X}_t\|_2^2 + \|\mathbb{E}\bar{X}_t - x\|_2^2 \right) \omega_\alpha(x, \mathbb{E}\bar{Y}_t) d\rho_{X,t}(x) \\ &\leq 2\text{Var}^X(t) + \frac{2}{\mathbb{E}\omega_\alpha(\bar{X}_t, \mathbb{E}\bar{Y}_t)} e^{-\alpha\mathcal{E}(\mathbb{E}\bar{Y}_t)} \text{Var}^X(t) \leq 4 \frac{\text{Var}^X(t)}{\mathcal{M}^X(t)} \end{aligned}$$

and analogously

$$(23) \quad \mathbb{E} \|\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2^2 \leq 4 \frac{\text{Var}^Y(t)}{\mathcal{M}^Y(t)},$$

which allow us to conclude the proof when being inserted into (18) and (19), respectively. \square

4.2. Time evolution of the functionals \mathcal{M}^X and \mathcal{M}^Y from (12b). In the time evolutions (17) of the variances Var^X and Var^Y , there appear the functionals \mathcal{M}^X and \mathcal{M}^Y as defined in (12b), which need to be controlled in order to ensure that the decay rates can be bounded from below by a positive constant, which eventually leads to at least exponential decay of the variances and therefore consensus of the dynamics (3). We therefore investigate the evolutions of \mathcal{M}^X and \mathcal{M}^Y next. To do so, let us recall from (12b) that $\mathcal{M}^X(t) = \widetilde{\mathcal{M}}^X(t) e^{\alpha\mathcal{E}(\mathbb{E}\bar{Y}_t)}$ and $\mathcal{M}^Y(t) = \widetilde{\mathcal{M}}^Y(t) e^{-\beta\mathcal{E}(\mathbb{E}\bar{X}_t)}$. We first bound in Lemma 14 the evolutions of $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ as defined in (12a), before we use product rule to obtain a lower bound for the evolutions of \mathcal{M}^X and \mathcal{M}^Y in Lemma 15. Let us furthermore remark that $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ will later allow us to characterize the convergence point of the dynamics (3).

LEMMA 14. *Let Var^X and Var^Y be as defined in (11) and $\widetilde{\mathcal{M}}^X$ and $\widetilde{\mathcal{M}}^Y$ as in (12a). Then, under assumptions A1 and A2, it holds that*

$$(24) \quad \frac{d}{dt} \widetilde{\mathcal{M}}^X(t) \geq -4\alpha e^{-\alpha\mathcal{E}(\mathbb{E}\bar{Y}_t)} C_{\nabla^2\mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(t)}{\mathcal{M}^X(t)} - \alpha\lambda_2 e^{-\alpha\mathcal{E}(\mathbb{E}\bar{Y}_t)} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}}$$

as well as

$$(25) \quad \frac{d}{dt} \widetilde{\mathcal{M}}^Y(t) \geq -4\beta e^{\beta\mathcal{E}(\mathbb{E}\bar{X}_t)} C_{\nabla^2\mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(t)}{\mathcal{M}^Y(t)} - \beta\lambda_1 e^{\beta\mathcal{E}(\mathbb{E}\bar{X}_t)} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^X(t)}}{\sqrt{\mathcal{M}^X(t)}}.$$

Proof. With Itô's formula and the chain rule, we first note that

$$\begin{aligned} d\widetilde{\mathcal{M}}^X(t) &= -\alpha \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) \nabla_x \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) \cdot d\bar{X}_t \right] \\ &\quad + \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) (\bar{X}_t - x_\alpha^Y(\rho_{X,t}))_k^2 \right. \\ &\quad \cdot \left. \left(\alpha^2 (\partial_{x_k} \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t))^2 - \alpha \partial_{x_k x_k}^2 \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) \right) \right] dt \\ &\quad - \alpha \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) \nabla_y \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) \cdot d\mathbb{E}\bar{Y}_t \right] =: (T_1 + T_2 + T_3) dt, \end{aligned}$$

where for the definition in the last step we exploited the fact that the appearing stochastic integrals have expectation 0 as a consequence of the regularity established in Theorem 6 and assumptions A1 and A2. Noticing that $\mathbb{E}[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t))(\bar{X}_t - x_\alpha^Y(\rho_{X,t}))] = 0$ and $\nabla_x \mathcal{E}(x_\alpha^Y(\rho_{X,t}), \mathbb{E}\bar{Y}_t)$ is deterministic, we obtain for T_1 the lower bound

$$\begin{aligned} T_1 &= \alpha\lambda_1 \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) \nabla_x \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) \cdot (\bar{X}_t - x_\alpha^Y(\rho_{X,t})) \right] \\ &= \alpha\lambda_1 \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) (\nabla_x \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) - \nabla_x \mathcal{E}(x_\alpha^Y(\rho_{X,t}), \mathbb{E}\bar{Y}_t)) \right. \\ &\quad \left. \cdot (\bar{X}_t - x_\alpha^Y(\rho_{X,t})) \right] \\ &\geq -\alpha\lambda_1 e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla^2 \mathcal{E}} \mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2, \end{aligned}$$

where we made use of the assumptions again. For T_2 , it holds that

$$\begin{aligned} T_2 &\geq -\alpha \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) (\bar{X}_t - x_\alpha^Y(\rho_{X,t}))_k^2 \partial_{x_k x_k}^2 \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) \right] dt \\ &\geq -\alpha \frac{\sigma_1^2}{2} e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla^2 \mathcal{E}} \mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2. \end{aligned}$$

And, eventually, for T_3 we have the following bound from below:

$$\begin{aligned} T_3 &= \alpha\lambda_2 \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t)) \nabla_y \mathcal{E}(\bar{X}_t, \mathbb{E}\bar{Y}_t) \cdot (\mathbb{E}\bar{Y}_t - y_\beta^X(\rho_{Y,t})) \right] \\ &\geq -\alpha\lambda_2 e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla \mathcal{E}} \|\mathbb{E}\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2, \end{aligned}$$

where we used the bounds on the gradient of \mathcal{E} required through assumption A1 in the last step. Collecting the estimates for T_1 , T_2 , and T_3 and inserting them into the first equation gives

$$\begin{aligned} \frac{d}{dt} \widetilde{\mathcal{M}}^X(t) &\geq -\alpha e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2 \\ &\quad - \alpha\lambda_2 e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla \mathcal{E}} \|\mathbb{E}\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2. \end{aligned}$$

The two appearing norms can be bounded by recalling (22) and noticing that (21) gives

$$(26) \quad \|\mathbb{E}\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2^2 \leq \frac{1}{\mathbb{E}\omega_{-\beta}(\mathbb{E}\bar{X}_t, \bar{Y}_t)} e^{\beta\xi(\mathbb{E}\bar{X}_t)} \int \|\mathbb{E}\bar{Y}_t - y\|_2^2 d\rho_{Y,t}(y) \leq \frac{\text{Var}^Y(t)}{\mathcal{M}^Y(t)}.$$

Inserting these two latter estimates allows us to continue the former as desired as

$$(27) \quad \frac{d}{dt} \widetilde{\mathcal{M}}^X(t) \geq -4\alpha e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(t)}{\mathcal{M}^X(t)} - \alpha\lambda_2 e^{-\alpha\xi(\mathbb{E}\bar{Y}_t)} C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}}.$$

The estimate for $\frac{d}{dt} \widetilde{\mathcal{M}}^Y(t)$ can be obtained analogously. \square

As mentioned already, we derive in the next lemma the time evolutions of the functionals \mathcal{M}^X and \mathcal{M}^Y as defined in (12b). This is an immediate consequence of product rule and Lemma 14.

LEMMA 15. Let Var^X and Var^Y be as defined in (11) and \mathcal{M}^X and \mathcal{M}^Y as in (12b). Then, under assumptions A1 and A2, it holds that

$$(28) \quad \frac{d}{dt} \mathcal{M}^X(t) \geq -4\alpha C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(t)}{\mathcal{M}^X(t)} - 2\alpha \lambda_2 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}}$$

as well as

$$(29) \quad \frac{d}{dt} \mathcal{M}^Y(t) \geq -4\beta C_{\nabla^2 \mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(t)}{\mathcal{M}^Y(t)} - 2\beta \lambda_1 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(t)}}{\sqrt{\mathcal{M}^X(t)}}.$$

Proof. By the product rule, we have $\frac{d}{dt} \mathcal{M}^X(t) = e^{\alpha \mathcal{E}(\mathbb{E}\bar{Y}_t)} \frac{d}{dt} \widetilde{\mathcal{M}}^X(t) + \widetilde{\mathcal{M}}^X(t) \frac{d}{dt} e^{\alpha \mathcal{E}(\mathbb{E}\bar{Y}_t)}$. While the first summand is controlled by recalling Lemma 14, for the second term we straightforwardly compute

$$\frac{d}{dt} e^{\alpha \mathcal{E}(\mathbb{E}\bar{Y}_t)} = \alpha e^{\alpha \mathcal{E}(\mathbb{E}\bar{Y}_t)} \nabla \mathcal{E}(\mathbb{E}\bar{Y}_t) \cdot \frac{d}{dt} \mathbb{E}\bar{Y}_t \geq -\alpha \lambda_2 e^{\alpha \mathcal{E}(\mathbb{E}\bar{Y}_t)} C_{\nabla \mathcal{E}} \|\mathbb{E}\bar{Y}_t - y_\beta^X(\rho_{Y,t})\|_2,$$

where we used the bounds on the gradient of \mathcal{E} required through assumption A1 together with the regularity from Theorem 6. Recalling (26) and putting everything together yields

$$\frac{d}{dt} \mathcal{M}^X(t) \geq -4\alpha C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(t)}{\mathcal{M}^X(t)} - (1 + \mathcal{M}^X(t)) \alpha \lambda_2 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}},$$

which gives the claim after noting that $\mathcal{M}^X(t) \leq 1$. The proceeding for $\frac{d}{dt} \mathcal{M}^Y(t)$ is identical. \square

4.3. Time evolution of the functionals \mathcal{M}_*^X and \mathcal{M}_*^Y from (12c). Similarly to the preceding sections, we study the time evolution of two functionals \mathcal{M}_*^X and \mathcal{M}_*^Y as defined in (12c), which aids in proving properties of the limit point of the mean-field dynamics (3).

LEMMA 16. Let Var^X and Var^Y be as defined in (11), \mathcal{M}^X and \mathcal{M}^Y as in (12b), and \mathcal{M}_*^X and \mathcal{M}_*^Y as in (12c). Then, under assumptions A1 and A2, it holds that

$$(30) \quad \frac{d}{dt} \mathcal{M}_*^X(t) \geq -4\alpha \lambda_1 e^{-\alpha \mathcal{E}(y^*)} C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(t)}}{\sqrt{\mathcal{M}^X(t)}} - 2\alpha \sigma_1^2 e^{-\alpha \mathcal{E}(y^*)} C_{\nabla^2 \mathcal{E}} \frac{\text{Var}^X(t)}{\mathcal{M}^X(t)}$$

as well as

$$(31) \quad \frac{d}{dt} \mathcal{M}_*^Y(t) \geq -4\beta \lambda_2 e^{\beta \bar{\mathcal{E}}(x^*)} C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(t)}}{\sqrt{\mathcal{M}^Y(t)}} - 2\beta \sigma_2^2 e^{\beta \bar{\mathcal{E}}(x^*)} C_{\nabla^2 \mathcal{E}} \frac{\text{Var}^Y(t)}{\mathcal{M}^Y(t)}.$$

Proof. With Itô's formula and the chain rule, we first note that

$$(32) \quad \begin{aligned} d\mathcal{M}_*^X(t) &= -\alpha \mathbb{E} \left[\exp(-\alpha \mathcal{E}(\bar{X}_t, y^*)) \nabla_x \mathcal{E}(\bar{X}_t, y^*) \cdot d\bar{X}_t \right] \\ &\quad + \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \mathbb{E} \left[\exp(-\alpha \mathcal{E}(\bar{X}_t, y^*)) (\bar{X}_t - x_\alpha^Y(\rho_{X,t}))_k^2 \right. \\ &\quad \left. \times \left(\alpha^2 (\partial_{x_k} \mathcal{E}(\bar{X}_t, y^*))^2 - \alpha \partial_{x_k x_k}^2 \mathcal{E}(\bar{X}_t, y^*) \right) \right] dt =: (T_1 + T_2) dt, \end{aligned}$$

where for the definition in the last step we again exploited the fact that the appearing stochastic integral has expectation 0 as a consequence of the assumptions. For T_1 , we have the lower bound

$$\begin{aligned} T_1 &\geq -\alpha\lambda_1 e^{-\alpha\xi(y^*)} \mathbb{E} \left[\|\nabla_x \mathcal{E}(\bar{X}_t, y^*)\|_2 \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2 \right] \\ &\geq -\alpha\lambda_1 e^{-\alpha\xi(y^*)} C_{\nabla\mathcal{E}} \sqrt{\mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2}. \end{aligned}$$

For T_2 , it holds that

$$\begin{aligned} T_2 &\geq -\alpha \frac{\sigma_1^2}{2} \sum_{k=1}^{d_1} \mathbb{E} \left[\exp(-\alpha\mathcal{E}(\bar{X}_t, y^*)) (\bar{X}_t - x_\alpha^Y(\rho_{X,t}))_k^2 \partial_{x_k x_k}^2 \mathcal{E}(\bar{X}_t, y^*) \right] dt \\ &\geq -\alpha \frac{\sigma_1^2}{2} e^{-\alpha\xi(y^*)} C_{\nabla^2\mathcal{E}} \mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2. \end{aligned}$$

Collecting the two former estimates for the terms T_1 and T_2 and inserting them into (32) gives

$$(33) \quad \begin{aligned} \frac{d}{dt} \mathcal{M}_*^X(t) &\geq -\alpha\lambda_1 e^{-\alpha\xi(y^*)} C_{\nabla\mathcal{E}} \sqrt{\mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2} \\ &\quad - \alpha \frac{\sigma_1^2}{2} e^{-\alpha\xi(y^*)} C_{\nabla^2\mathcal{E}} \mathbb{E} \|\bar{X}_t - x_\alpha^Y(\rho_{X,t})\|_2^2, \end{aligned}$$

where the last expression can be bounded by employing (22). The estimate for $\frac{d}{dt} \mathcal{M}_*^Y(t)$ can be obtained analogously. \square

4.4. Proof of Theorem 11.

Proof of Theorem 11.

Step 1a. Let us define the time horizon

$$(34) \quad T := \inf \left\{ t \geq 0 : \mathcal{M}^X(t) < \frac{1}{2} \mathcal{M}^X(0) \text{ or } \mathcal{M}^Y(t) < \frac{1}{2} \mathcal{M}^Y(0) \right\} \quad \text{with } \inf \emptyset = \infty,$$

where \mathcal{M}^X and \mathcal{M}^Y are as defined in (12b). Obviously, by continuity, $T > 0$. We claim that $T = \infty$, which is shown by contradiction in what follows. Therefore, let us assume $T < \infty$. Then, as a consequence of the definition of the time horizon T , the prefactors of $\text{Var}^X(t)$ and $\text{Var}^Y(t)$ in Lemma 13 are upper bounded by $-\mu_1$ and $-\mu_2$, respectively, for all $t \in [0, T]$. Consequently, Lemma 13 permits the upper bounds $\frac{d}{dt} \text{Var}^X(t) \leq -\mu_1 \text{Var}^X(t)$ and $\frac{d}{dt} \text{Var}^Y(t) \leq -\mu_2 \text{Var}^Y(t)$ for the time evolution of the functionals Var^X and Var^Y . The negativity of the rate is ensured by the well-preparedness condition P1. An application of Grönwall's inequality gives

$$(35) \quad \text{Var}^X(t) \leq \text{Var}^X(0) e^{-\mu_1 t} \quad \text{and} \quad \text{Var}^Y(t) \leq \text{Var}^Y(0) e^{-\mu_2 t}.$$

Let us now derive the contradiction. It follows from Lemma 15 for \mathcal{M}^X and \mathcal{M}^Y from (12b) that

$$(36) \quad \begin{aligned} \frac{d}{dt} \mathcal{M}^X(t) &\geq -8\alpha C_{\nabla^2\mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(0) e^{-\mu_1 t}}{\mathcal{M}^X(0)} - 2\sqrt{2}\alpha\lambda_2 C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^Y(0) e^{-\mu_2 t/2}}}{\sqrt{\mathcal{M}^Y(0)}}, \\ \frac{d}{dt} \mathcal{M}^Y(t) &\geq -8\beta C_{\nabla^2\mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(0) e^{-\mu_2 t}}{\mathcal{M}^Y(0)} - 2\sqrt{2}\beta\lambda_1 C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^X(0) e^{-\mu_1 t/2}}}{\sqrt{\mathcal{M}^X(0)}}, \end{aligned}$$

where we used the formerly derived (35) as well as the fact that $\mathcal{M}^X(t) \geq \mathcal{M}^X(0)/2$ and $\mathcal{M}^Y(t) \geq \mathcal{M}^Y(0)/2$ for all $t \in [0, T]$ by the definition of T . Integrating (36) and employing the well-preparedness condition P3 shows that for all $t \in [0, T]$,

$$\begin{aligned} \mathcal{M}^X(t) &\geq \mathcal{M}^X(0) - 8\alpha C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(0)}{\mu_1 \mathcal{M}^X(0)} - 4\sqrt{2}\alpha\lambda_2 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)}}{\mu_2 \sqrt{\mathcal{M}^Y(0)}} \\ &\geq \frac{3}{4} \mathcal{M}^X(0), \\ \mathcal{M}^Y(t) &\geq \mathcal{M}^Y(0) - 8\beta C_{\nabla^2 \mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(0)}{\mu_2 \mathcal{M}^Y(0)} - 4\sqrt{2}\beta\lambda_1 C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(0)}}{\mu_1 \sqrt{\mathcal{M}^X(0)}} \\ &\geq \frac{3}{4} \mathcal{M}^Y(0). \end{aligned}$$

This entails the fact that there exists $\delta > 0$ such that $\mathcal{M}^X(t) \geq \mathcal{M}^X(0)/2$ and $\mathcal{M}^Y(t) \geq \mathcal{M}^Y(0)/2$ hold for all $t \in [T, T + \delta]$ as well, contradicting the definition of T and therefore showing that $T = \infty$. Consequently, (35) as well as

$$(37) \quad \mathcal{M}^X(t) \geq \frac{1}{2} \mathcal{M}^X(0) \quad \text{and} \quad \mathcal{M}^Y(t) \geq \frac{1}{2} \mathcal{M}^Y(0)$$

hold for all $t \geq 0$, which proves (13).

Step 1b. With Jensen's inequality and by making use of the bounds (22) and (23) combined with (35) and (37), we further observe that

$$\begin{aligned} \left\| \frac{d}{dt} \mathbb{E} \bar{X}_t \right\|_2 &\leq \lambda_1 \mathbb{E} \left\| \bar{X}_t - x_\alpha^Y(\rho_{X,t}) \right\|_2 \leq 2\lambda_1 \frac{\sqrt{\text{Var}^X(0)} e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}} \rightarrow 0 \quad \text{as } t \rightarrow \infty, \\ \left\| \frac{d}{dt} \mathbb{E} \bar{Y}_t \right\|_2 &\leq \lambda_2 \mathbb{E} \left\| \bar{Y}_t - y_\beta^X(\rho_{Y,t}) \right\|_2 \leq 2\lambda_2 \frac{\sqrt{\text{Var}^Y(0)} e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}} \rightarrow 0 \quad \text{as } t \rightarrow \infty. \end{aligned}$$

We therefore have $(\mathbb{E} \bar{X}_t, \mathbb{E} \bar{Y}_t) \rightarrow (\tilde{x}, \tilde{y})$ for some $(\tilde{x}, \tilde{y}) \in \mathbb{R}^{d_1+d_2}$. In fact, following from (35), $(\bar{X}_t, \bar{Y}_t) \rightarrow (\tilde{x}, \tilde{y})$ and $(x_\alpha^Y(\rho_{X,t}), y_\beta^X(\rho_{Y,t})) \rightarrow (\tilde{x}, \tilde{y})$ in L^2 thanks to (22) and (23). This shows (14).

Step 2a. It remains to verify (15) for the point (\tilde{x}, \tilde{y}) . With arguments similar to those in Step 1a, let us first derive analogous statements as in (37) for $\tilde{\mathcal{M}}^X$ and $\tilde{\mathcal{M}}^Y$ as defined in (12a) as well as \mathcal{M}_*^X and \mathcal{M}_*^Y as defined in (12c), respectively. To do so, we first notice that $(\mathbb{E} \bar{X}_t, \mathbb{E} \bar{Y}_t)$ is continuous and since it converges to (\tilde{x}, \tilde{y}) as $t \rightarrow \infty$, there exists $M > 0$, potentially depending on (\tilde{x}, \tilde{y}) , such that $\|\mathbb{E} \bar{X}_t\|_2 + \|\mathbb{E} \bar{Y}_t\|_2 \leq M$ for all $t \geq 0$. Since, moreover, \mathcal{E} and $\bar{\mathcal{E}}$ are continuous, there exists $\mathcal{E}_M > 0$ such that $-\mathcal{E}_M \leq \mathcal{E}(\mathbb{E} \bar{Y}_t) \leq \bar{\mathcal{E}}(\mathbb{E} \bar{X}_t) \leq \mathcal{E}_M$ for all $t > 0$. Utilizing this together with (35) and (37), we derive from Lemma 14 for $\tilde{\mathcal{M}}^X$ and $\tilde{\mathcal{M}}^Y$ from (12a) that

$$\begin{aligned} \frac{d}{dt} \tilde{\mathcal{M}}^X(t) &\geq -8\alpha e^{\alpha \mathcal{E}_M} C_{\nabla^2 \mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(0) e^{-\mu_1 t}}{\mathcal{M}^X(0)} \\ &\quad - \sqrt{2}\alpha\lambda_2 e^{\alpha \mathcal{E}_M} C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)} e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}}, \\ (38) \quad \frac{d}{dt} \tilde{\mathcal{M}}^Y(t) &\geq -8\beta e^{\beta \mathcal{E}_M} C_{\nabla^2 \mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(0) e^{-\mu_2 t}}{\mathcal{M}^Y(0)} \\ &\quad - \sqrt{2}\beta\lambda_1 e^{\beta \mathcal{E}_M} C_{\nabla \mathcal{E}} \frac{\sqrt{\text{Var}^X(0)} e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}}. \end{aligned}$$

Analogously, by using (35) and (37) it follows directly from Lemma 16 for \mathcal{M}_*^X and \mathcal{M}_*^Y from (12c) that

$$\begin{aligned}
 \frac{d}{dt} \mathcal{M}_*^X(t) &\geq -8\alpha\lambda_1 e^{-\alpha\xi(y^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^X(0)} e^{-\mu_1 t/2}}{\sqrt{\mathcal{M}^X(0)}} \\
 &\quad - 4\alpha\sigma_1^2 e^{-\alpha\xi(y^*)} C_{\nabla^2\mathcal{E}} \frac{\text{Var}^X(0) e^{-\mu_1 t}}{\mathcal{M}^X(0)}, \\
 \frac{d}{dt} \mathcal{M}_*^Y(t) &\geq -8\beta\lambda_2 e^{\beta\bar{\xi}(x^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)} e^{-\mu_2 t/2}}{\sqrt{\mathcal{M}^Y(0)}} \\
 &\quad - 4\beta\sigma_2^2 e^{\beta\bar{\xi}(x^*)} C_{\nabla^2\mathcal{E}} \frac{\text{Var}^Y(0) e^{-\mu_2 t}}{\mathcal{M}^Y(0)}.
 \end{aligned}
 \tag{39}$$

Integrating (38) and employing the well-preparedness condition P3 shows for all $t \geq 0$ that

$$\begin{aligned}
 \widetilde{\mathcal{M}}^X(t) &\geq \widetilde{\mathcal{M}}^X(0) - 8\alpha e^{\alpha\mathcal{E}_M} C_{\nabla^2\mathcal{E}} \left(\lambda_1 + \frac{\sigma_1^2}{2} \right) \frac{\text{Var}^X(0)}{\mu_1 \mathcal{M}^X(0)} \\
 &\quad - 2\sqrt{2}\alpha\lambda_2 e^{\alpha\mathcal{E}_M} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)}}{\mu_2 \sqrt{\mathcal{M}^Y(0)}} \geq \frac{3}{4} \widetilde{\mathcal{M}}^X(0), \\
 \widetilde{\mathcal{M}}^Y(t) &\geq \widetilde{\mathcal{M}}^Y(0) - 8\beta e^{\beta\mathcal{E}_M} C_{\nabla^2\mathcal{E}} \left(\lambda_2 + \frac{\sigma_2^2}{2} \right) \frac{\text{Var}^Y(0)}{\mu_2 \mathcal{M}^Y(0)} \\
 &\quad - 2\sqrt{2}\beta\lambda_1 e^{\beta\mathcal{E}_M} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^X(0)}}{\mu_1 \sqrt{\mathcal{M}^X(0)}} \geq \frac{3}{4} \widetilde{\mathcal{M}}^Y(0).
 \end{aligned}$$

Analogously, using (39) together with P3 shows for all $t \geq 0$ that

$$\begin{aligned}
 \mathcal{M}_*^X(t) &\geq \mathcal{M}_*^X(0) - 16\alpha\lambda_1 e^{-\alpha\xi(y^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^X(0)}}{\mu_1 \sqrt{\mathcal{M}^X(0)}} \\
 &\quad - 4\alpha\sigma_1^2 e^{-\alpha\xi(y^*)} C_{\nabla^2\mathcal{E}} \frac{\text{Var}^X(0)}{\mu_1 \mathcal{M}^X(0)} \geq \frac{3}{4} \mathcal{M}_*^X(0), \\
 \mathcal{M}_*^Y(t) &\geq \mathcal{M}_*^Y(0) - 16\beta\lambda_2 e^{\beta\bar{\xi}(x^*)} C_{\nabla\mathcal{E}} \frac{\sqrt{\text{Var}^Y(0)}}{\mu_2 \sqrt{\mathcal{M}^Y(0)}} \\
 &\quad - 4\beta\sigma_2^2 e^{\beta\bar{\xi}(x^*)} C_{\nabla^2\mathcal{E}} \frac{\text{Var}^Y(0)}{\mu_2 \mathcal{M}^Y(0)} \geq \frac{3}{4} \mathcal{M}_*^Y(0).
 \end{aligned}$$

Thus, in particular it holds that for all $t \geq 0$,

$$\widetilde{\mathcal{M}}^X(t) \geq \frac{1}{2} \widetilde{\mathcal{M}}^X(0) \quad \text{and} \quad \widetilde{\mathcal{M}}^Y(t) \geq \frac{1}{2} \widetilde{\mathcal{M}}^Y(0)
 \tag{40}$$

as well as

$$\mathcal{M}_*^X(t) \geq \frac{1}{2} \mathcal{M}_*^X(0) \quad \text{and} \quad \mathcal{M}_*^Y(t) \geq \frac{1}{2} \mathcal{M}_*^Y(0).
 \tag{41}$$

Step 2b. By Chebyshev’s inequality, for each $\delta > 0$ it holds that

$$\rho_t(\{\|x - \tilde{x}, y - \tilde{y}\|_2 \geq \delta\}) \leq \frac{2}{\delta^2} (\text{Var}^X(t) + \text{Var}^Y(t) + \|\mathbb{E}\bar{X}_t - \tilde{x}\|_2^2 + \|\mathbb{E}\bar{Y}_t - \tilde{y}\|_2^2),$$

which converges to 0 as $t \rightarrow \infty$. Thus, the pair (\bar{X}_t, \bar{Y}_t) converges to (\tilde{x}, \tilde{y}) in probability as $t \rightarrow \infty$. Recall the convergence $(\mathbb{E}\bar{X}_t, \mathbb{E}\bar{Y}_t) \rightarrow (\tilde{x}, \tilde{y})$, the continuity of \mathcal{E} , and the fact that for all $t \geq 0$, $\exp(-\alpha\mathcal{E}(\bar{X}_t, \bar{Y}_t)) \leq \exp(\alpha\mathcal{E}_M)$ holds a.s.

By the dominated convergence theorem, one can pass to the limit in t to obtain $\lim_{t \rightarrow \infty} \widetilde{\mathcal{M}}^X(t) = \exp(-\alpha \mathcal{E}(\widetilde{x}, \widetilde{y}))$. Analogously, one may get $\widetilde{\mathcal{M}}^Y(t) \rightarrow \exp(\beta \mathcal{E}(\widetilde{x}, \widetilde{y}))$ as $t \rightarrow \infty$. Using this when taking $t \rightarrow \infty$ in the bounds (40) after applying the logarithm and multiplying both sides with $-1/\alpha$ and $1/\beta$, respectively, we obtain

$$(42) \quad \begin{aligned} \mathcal{E}(\widetilde{x}, \widetilde{y}) &= \lim_{t \rightarrow \infty} \left(-\frac{1}{\alpha} \log \widetilde{\mathcal{M}}^X(t) \right) \leq \frac{1}{\alpha} \log 2 - \frac{1}{\alpha} \log \widetilde{\mathcal{M}}^X(0), \\ \mathcal{E}(\widetilde{x}, \widetilde{y}) &= \lim_{t \rightarrow \infty} \left(\frac{1}{\beta} \log \widetilde{\mathcal{M}}^Y(t) \right) \geq -\frac{1}{\beta} \log 2 + \frac{1}{\beta} \log \widetilde{\mathcal{M}}^Y(0). \end{aligned}$$

Due to the first set of well-preparedness conditions from P2, the Laplace principle in the form of Lemmas A.3 and A.4 when choosing μ^α as the law of the initial data \overline{X}_0 and μ^β as the law of \overline{Y}_0 now allows one to choose $\alpha \geq (2 \log 2)/\varepsilon$ and $\beta \geq (2 \log 2)/\varepsilon$ large enough such that for given $\varepsilon > 0$ it moreover holds that

$$\begin{aligned} -\frac{1}{\alpha} \log \widetilde{\mathcal{M}}^X(0) - \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}\overline{Y}_0) &= -\frac{1}{\alpha} \log \mathbb{E} \exp(-\alpha \mathcal{E}(\overline{X}_0, \mathbb{E}\overline{Y}_0)) - \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}\overline{Y}_0) \\ &\leq \varepsilon/2, \\ -\frac{1}{\beta} \log \widetilde{\mathcal{M}}^Y(0) + \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}\overline{X}_0, y) &= -\frac{1}{\beta} \log \mathbb{E} \exp(\beta \mathcal{E}(\mathbb{E}\overline{X}_0, \overline{Y}_0)) + \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}\overline{X}_0, y) \\ &\leq \varepsilon/2. \end{aligned}$$

Notice here that we well-prepare α and β simultaneously with the initial data $(\overline{X}_0, \overline{Y}_0)$ (therewith $(\overline{X}_0, \overline{Y}_0)$ depends on α, β). However, due to the well-preparedness conditions P2, α and β can still be taken sufficiently large as ensured in Lemmas A.3 and A.4.

Such choices of parameters in (42) immediately give $\mathcal{E}(\widetilde{x}, \widetilde{y}) \leq \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, \mathbb{E}\overline{Y}_0) + \varepsilon$ and $\mathcal{E}(\widetilde{x}, \widetilde{y}) \geq \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(\mathbb{E}\overline{X}_0, y) - \varepsilon$ and consequently $\mathcal{E}(\widetilde{x}, \widetilde{y}) \leq \min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x, y) + \varepsilon$ and $\mathcal{E}(\widetilde{x}, \widetilde{y}) \geq \max_{y \in \mathbb{R}^{d_2}} \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y) - \varepsilon$, which proves the first part of (15). Second, following an analogous argumentation for \mathcal{M}_*^X and \mathcal{M}_*^Y as defined in (12c), we obtain the remainder of (15). More precisely, we first note that $\mathcal{M}_*^X(t) \rightarrow \exp(-\alpha \mathcal{E}(\widetilde{x}, y^*))$ and $\mathcal{M}_*^Y(t) \rightarrow \exp(\beta \mathcal{E}(x^*, \widetilde{y}))$ as $t \rightarrow \infty$. Taking now the limit $t \rightarrow \infty$ in (41) after suitable algebraic manipulations, we obtain

$$(43) \quad \begin{aligned} \mathcal{E}(\widetilde{x}, y^*) &= \lim_{t \rightarrow \infty} \left(-\frac{1}{\alpha} \log \mathcal{M}_*^X(t) \right) \leq \frac{1}{\alpha} \log 2 - \frac{1}{\alpha} \log \mathcal{M}_*^X(0), \\ \mathcal{E}(x^*, \widetilde{y}) &= \lim_{t \rightarrow \infty} \left(\frac{1}{\beta} \log \mathcal{M}_*^Y(t) \right) \geq -\frac{1}{\beta} \log 2 + \frac{1}{\beta} \log \mathcal{M}_*^Y(0). \end{aligned}$$

A potentially larger choice of α and β allows (again by the Laplace principle in the form of Lemmas A.3 and A.4, which applies due to the second set of well-preparedness conditions from P2) us to guarantee

$$\begin{aligned} -\frac{1}{\alpha} \log \mathcal{M}_*^X(0) - \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y^*) &= -\frac{1}{\alpha} \log \mathbb{E} \exp(-\alpha \mathcal{E}(\overline{X}_0, y^*)) - \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y^*) \leq \varepsilon/2, \\ -\frac{1}{\beta} \log \mathcal{M}_*^Y(0) + \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x^*, y) &= -\frac{1}{\beta} \log \mathbb{E} \exp(\beta \mathcal{E}(x^*, \overline{Y}_0)) + \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x^*, y) \leq \varepsilon/2 \end{aligned}$$

for the specified ε . Such choices of parameters in (43) immediately give $\mathcal{E}(\widetilde{x}, y^*) \leq \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y^*) + \varepsilon$ and $\mathcal{E}(x^*, \widetilde{y}) \geq \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x^*, y) - \varepsilon$, which completes the proof of (15).

Step 3. Finally, under the inverse continuity property A3 and making use of what we just proved, we additionally obtain (16), which concludes the proof. \square

5. Implementation of CBO-SP and numerical experiments.

5.1. Numerical algorithm and implementation. In order to implement and run CBO-SP on a computer, we first fix a discrete time step size Δt and a number of iterations K or define any other suitable stopping criterion. Then, by discretizing the interacting particle system (1) via an Euler–Maruyama time discretization [19, 41] as

$$(44a) \quad \widehat{X}_{k+1}^i = \widehat{X}_k^i - \lambda_1 \Delta t \left(\widehat{X}_k^i - x_\alpha^Y(\widehat{\rho}_{X,k}^{N_1}) \right) + \sigma_1 D \left(\widehat{X}_k^i - x_\alpha^Y(\widehat{\rho}_{X,k}^{N_1}) \right) B_k^{X,i},$$

$$(44b) \quad \widehat{Y}_{k+1}^i = \widehat{Y}_k^i - \lambda_2 \Delta t \left(\widehat{Y}_k^i - y_\beta^X(\widehat{\rho}_{Y,k}^{N_2}) \right) + \sigma_2 D \left(\widehat{Y}_k^i - y_\beta^X(\widehat{\rho}_{Y,k}^{N_2}) \right) B_k^{Y,i},$$

where $\widehat{\rho}_{X,k}^{N_1}$ and $\widehat{\rho}_{Y,k}^{N_2}$ denote the empirical averages of $(\widehat{X}_k^i)_{i=1,\dots,N_1}$ and $(\widehat{Y}_k^i)_{i=1,\dots,N_2}$ and where

$$(45a) \quad \widehat{x}_\alpha^Y(\widehat{\rho}_{X,k}^{N_1}) = \int x \frac{\omega_\alpha(x, \int y d\widehat{\rho}_{Y,k}^{N_2}(y))}{\|\omega_\alpha(\cdot, \int y d\widehat{\rho}_{Y,k}^{N_2}(y))\|_{L^1(\widehat{\rho}_{X,k}^{N_1})}} d\widehat{\rho}_{X,k}^{N_1}(x),$$

$$(45b) \quad \widehat{y}_\beta^X(\widehat{\rho}_{Y,k}^{N_2}) = \int y \frac{\omega_{-\beta}(\int x d\widehat{\rho}_{X,k+1}^{N_1}(x), y)}{\|\omega_{-\beta}(\int x d\widehat{\rho}_{X,k+1}^{N_1}(x), \cdot)\|_{L^1(\widehat{\rho}_{Y,k}^{N_2})}} d\widehat{\rho}_{Y,k}^{N_2}(y),$$

we obtain the implementable iterative scheme, which is used in the formulation of Algorithm 1. Moreover, $((B_k^{X,i})_{k=1,\dots,K})_{i=1,\dots,N_1}$ and $((B_k^{Y,i})_{k=1,\dots,K})_{i=1,\dots,N_2}$ in (44) are independent Gaussian vectors in \mathbb{R}^{d_1} and \mathbb{R}^{d_2} , respectively, with covariance matrix $\Delta t I_d$. Note that in (45b) we could also use the old iterates $\widehat{\rho}_{X,k}^{N_1}$ instead of the new ones $\widehat{\rho}_{X,k+1}^{N_1}$ for the computation.

5.2. Illustrative numerical experiments for CBO-SP. To visualize the behavior of the CBO-SP algorithm in practice, we depict in Figure 1 snapshots of the positions of the particles for four different types of saddle point functions, which are plotted in the first row of the figure. The experiments include two *nonconvex-nonconcave* examples, which is in general the setting of particular interest in modern

Algorithm 1. CBO-SP.

Input: Objective \mathcal{E} , discrete time step size Δt , number of iterates K , parameters $\lambda_1, \lambda_2, \sigma_1, \sigma_2, \alpha, \beta$, number of particles N_1 and N_2 , initialization ρ_0

Output: Approximation $(\widehat{x}_\alpha^Y(\widehat{\rho}_{X,k}^{N_1}), \widehat{y}_\beta^X(\widehat{\rho}_{Y,k}^{N_2}))$ of the saddle point (x^*, y^*) of \mathcal{E}

1: Generate the particles' initial positions $(X_0^i)_{i=1,\dots,N_1}$ and $(Y_0^i)_{i=1,\dots,N_2}$

according to the initial laws $\rho_{X,0}$ and $\rho_{Y,0}$, respectively. Set $k = 0$.

2: **while** $k \leq K$ or stopping criterion not fulfilled

3: Compute the component $\widehat{x}_\alpha^Y(\widehat{\rho}_{X,k}^{N_1})$ of the consensus point according to (45a).

4: Update the X -positions by computing $(\widehat{X}_{k+1}^i)_{i=1,\dots,N_1}$ according to (44a).

5: Compute the component $\widehat{y}_\beta^X(\widehat{\rho}_{Y,k}^{N_2})$ of the consensus point according to (45b).

6: Update the Y -positions by computing $(\widehat{Y}_{k+1}^i)_{i=1,\dots,N_2}$ according to (44b).

7: Check the stopping criterion and **break** if fulfilled. If not, continue and set $k = k + 1$.

8: **end while**

9: Compute consensus point $(\widehat{x}_\alpha^Y(\widehat{\rho}_{X,k}^{N_1}), \widehat{y}_\beta^X(\widehat{\rho}_{Y,k}^{N_2}))$ as final approximation to saddle point (x^*, y^*) .

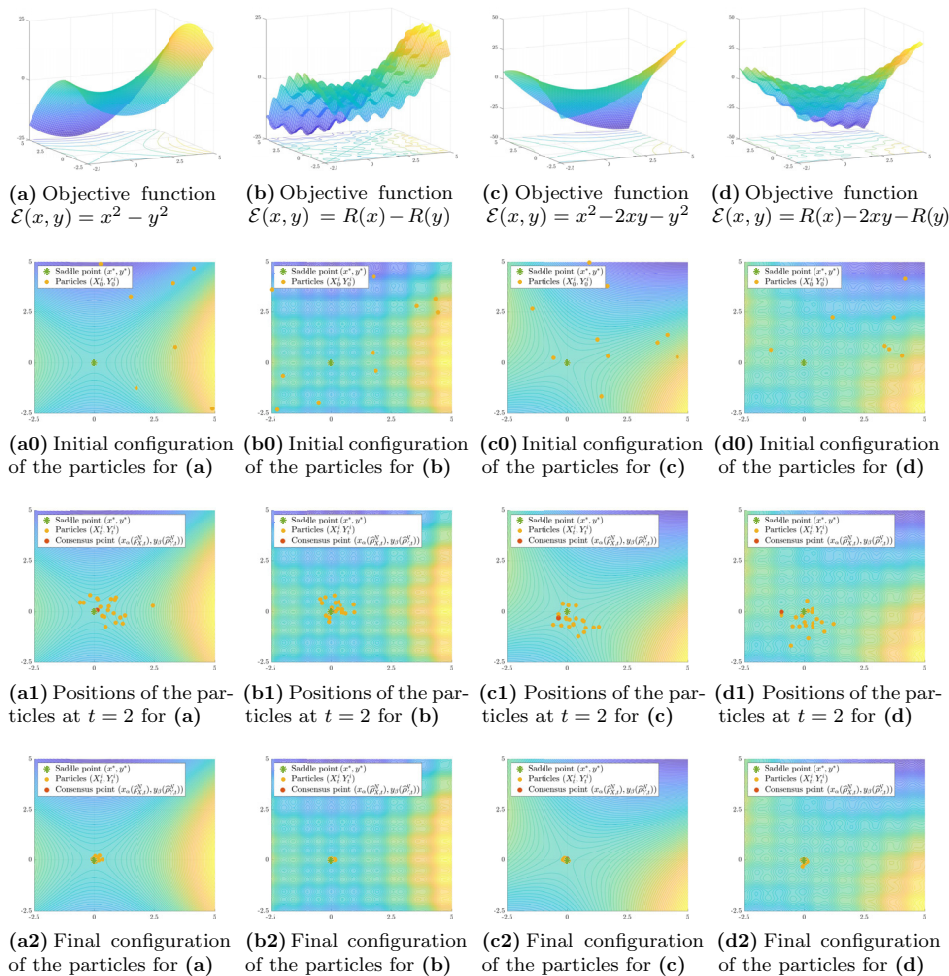


FIG. 1. Illustration of the dynamics of CBO-SP when searching the global Nash equilibrium of four different saddle point functions plotted in (a)–(d) in $d = 1$, where $R(x) = \sum_{k=1}^d x_k^2 + \frac{5}{2}(1 - \cos(2\pi x_k))$ is the Rastrigin function. Each column visualizes the positions of the $N = 20$ particles when running CBO-SP with parameters $\alpha = \beta = 10^{15}$, $\lambda_1 = \lambda_2 = 1$, $\sigma_1 = \sigma_2 = \sqrt{0.1}$ and time step size $\Delta t = 0.1$ at three different points in time ($t = 0$, $t = 2$, and $t = T = 4$). The particles are sampled initially from $\rho_0 \sim N(2, 4) \times N(2, 4)$.

applications. We observe that in all cases (also in case of different initializations) the saddle point is found fast and reliably.

5.3. Solving a quadratic game with CBO-SP. To demonstrate the practicability of CBO-SP, we solve a strongly monotone quadratic game [27, section 5] of the form $\min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathbb{R}^{d_2}} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} x^T A_i x + x^T B_i y - \frac{1}{2} x^T C_i y$ with sample size $n = 100$ and for various dimensions d_1 and d_2 . The matrices $B_i \in \mathbb{R}^{d_1 \times d_2}$ have random Gaussian entries, and the positive definite matrices $A_i \in \mathbb{R}^{d_1 \times d_1}$ and $C_i \in \mathbb{R}^{d_2 \times d_2}$ are of the form $A_i = \tilde{A}_i^T \tilde{A}_i$ and $C_i = \tilde{C}_i^T \tilde{C}_i$ with $\tilde{A}_i \in \mathbb{R}^{d_1 \times d_1}$ and $\tilde{C}_i \in \mathbb{R}^{d_2 \times d_2}$

TABLE 1

Success rates, average runtime (in ms), and average ℓ^∞ -error of the CBO-SP algorithm when solving a quadratic game for different dimensions d_1 and d_2 and with different numbers of particles N . All results are computed on the basis of 100 runs of the algorithm.

	$N = 40$	$N = 80$	$N = 120$	$N = 200$
$d_1 = 20, d_2 = 8$	31% (32ms, $1.5 \cdot 10^{-2}$)	100% (50ms, $2.4 \cdot 10^{-7}$)	100% (182ms, $6.1 \cdot 10^{-8}$)	100% (229ms, $2.9 \cdot 10^{-8}$)
$d_1 = 20, d_2 = 20$	7% (41ms, $3.0 \cdot 10^{-2}$)	100% (63ms, $4.5 \cdot 10^{-7}$)	100% (339ms, $3.7 \cdot 10^{-8}$)	100% (462ms, $2.4 \cdot 10^{-8}$)
$d_1 = 40, d_2 = 8$	0% (154ms, 1.1)	1% (190ms, $3.6 \cdot 10^{-2}$)	53% (226ms, $2.6 \cdot 10^{-3}$)	100% (418ms, $4.8 \cdot 10^{-5}$)
$d_1 = 40, d_2 = 20$	0% (162ms, 1.2)	0% (285ms, $4.9 \cdot 10^{-2}$)	52% (436ms, $3.8 \cdot 10^{-3}$)	100% (539ms, $8.2 \cdot 10^{-5}$)
$d_1 = 40, d_2 = 40$	0% (330ms, 1.9)	0% (336ms, $1.2 \cdot 10^{-1}$)	25% (421ms, $5.4 \cdot 10^{-3}$)	100% (606ms, $7.9 \cdot 10^{-5}$)

having random Gaussian entries. We employ CBO-SP with parameters $\alpha = \beta = 10^{15}$, $\lambda_1 = \lambda_2 = 1$, $\sigma_1 = \sigma_2 = 2$ using $N \in \{40, 80, 120, 200\}$ particles and with time horizon $T = 100$ and discrete time step size $\Delta t = 0.1$. The particles are sampled initially from $\rho_0 \sim \mathcal{N}(4, 2\text{Id}) \times \mathcal{N}(4, 2\text{Id})$ (i.e., they are initialized substantially far from the saddle point). We depict in Table 1 the success rates, average ℓ^∞ -error, and average runtime of the CBO-SP algorithm computed on the basis of 100 runs. A run is considered successful if the obtained solution has an accuracy of 10^{-3} w.r.t. the ℓ^∞ -norm. In brackets, we indicate the average (over the runs) runtime in milliseconds (ms) as well as the average (over the runs) ℓ^∞ -error.

We observe that with the already moderately many particles, the CBO-SP algorithm is capable of consistently finding the desired saddle point for relatively high-dimensional minimax problems.

Experiments in much higher dimensions and more applied settings coming, for instance, from economics or arising when training GANs are left to future and more experimental research, which focuses on benchmarking rather than providing rigorous convergence guarantees.

6. Conclusions. In this paper, we propose consensus-based optimization for saddle point problems (CBO-SP) and analyze its global convergence behavior to global Nash equilibria. As is apparent from the proof, our technique requires the equilibrium to satisfy the saddle point property, i.e., that $\min_x \max_y$ and $\max_y \min_x$ coincide. We leave to further research the extension of the results to sequential games, where the latter condition does not hold. This is in particular relevant in, for instance, the training of GANs, which are formulated as nonsimultaneous games.

Appendix A.

A.1. Existence and uniqueness of solutions to SDEs. For the sake of self-consistency, we recall two results from [10] about the existence and pathwise uniqueness of a strong solution of an SDE of the form $Z_t = Z_0 + \int_0^t b(Z_s) ds + \int_0^t \sigma(Z_s) dB_s$ (\star). These results are used in the proof of Theorem 3. Note that here we adopted the notation of [10]; i.e., in our setting, we have $Z_t = \mathbf{Z}_t$ as well as $b(\mathbf{Z}_t) = -\lambda \mathbf{F}(\mathbf{Z}_t)$ and $\sigma(\mathbf{Z}_t) = \sigma \mathbf{M}(\mathbf{Z}_t)$.

THEOREM A.1 (see [10, Chapter 5, Theorem 3.1]). *Suppose the following:*

- (i) *For any $n < \infty$, we have $|b_i(z) - b_i(z')| \leq K_n |z - z'|$ and $|\sigma_{ij}(z) - \sigma_{ij}(z')| \leq K_n |z - z'|$ for $|z|, |z'| \leq n$.*
- (ii) *There exist a constant $A < \infty$ and a function $\varphi(z) \geq 0$ so that if Z_t is a solution of (\star), then $e^{-At} \varphi(Z_t)$ is a local supermartingale.*

Then (\star) has a strong solution and pathwise uniqueness holds.

THEOREM A.2 (see [10, Chapter 5, Theorem 3.2]). Let $a = \sigma\sigma^T$, and suppose that $\sum_{i=1}^d 2z_i b_i(z) + a_{ii}(z) \leq B(1 + |z|^2)$. Then (ii) in Theorem A.1 holds with $A = B$ and $\varphi(z) = 1 + |z|^2$.

A.2. The Laplace principle.

LEMMA A.3. Define $S_{y,\delta} = \{x \in \mathbb{R}^{d_1} : \exp(-\mathcal{E}(x, y)) > \exp(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y)) - \delta\}$ for any fixed $y \in \mathbb{R}^{d_2}$ and any $\delta > 0$. Let $\{\mu^\alpha\}_{\alpha \geq 1}$ be a family of measures in $\mathcal{P}(\mathbb{R}^{d_1})$, and assume there exists a constant $C_\delta > 0$ depending only on δ such that $\mu^\alpha(S_{y,\delta}) \geq C_\delta$ for all $\alpha \geq 1$. Then it holds that

$$\lim_{\alpha \rightarrow \infty} -\frac{1}{\alpha} \log \left(\int_{\mathbb{R}^{d_1}} \exp(-\alpha \mathcal{E}(x, y)) d\mu^\alpha(x) \right) = \min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y).$$

Proof. We first notice that by the definition of the set $S_{y,\delta}$ it holds that

$$\begin{aligned} \left(\int_{\mathbb{R}^{d_1}} \exp(-\alpha \mathcal{E}(x, y)) d\mu^\alpha(x) \right)^{1/\alpha} &\geq \left(\int_{S_{y,\delta}} \left(\exp\left(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y)\right) - \delta \right)^\alpha d\mu^\alpha(x) \right)^{1/\alpha} \\ &= \left(\exp\left(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y)\right) - \delta \right) \mu^\alpha(S_{y,\delta})^{1/\alpha} \\ &\geq \left(\exp\left(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y)\right) - \delta \right) C_\delta^{1/\alpha}, \end{aligned}$$

which converges to $\exp(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y)) - \delta$ as $\alpha \rightarrow \infty$. Thus, for any $\delta > 0$, we have $\liminf_{\alpha \rightarrow \infty} \left(\int_{\mathbb{R}^{d_1}} \exp(-\alpha \mathcal{E}(x, y)) d\mu^\alpha(x) \right)^{1/\alpha} \geq \exp(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y)) - \delta$. On the other hand, clearly $\limsup_{\alpha \rightarrow \infty} \left(\int_{\mathbb{R}^{d_1}} \exp(-\alpha \mathcal{E}(x, y)) d\mu^\alpha(x) \right)^{1/\alpha} \leq \exp(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y))$. Since δ was arbitrary, this implies that $\lim_{\alpha \rightarrow \infty} \left(\int_{\mathbb{R}^{d_1}} \exp(-\alpha \mathcal{E}(x, y)) d\mu^\alpha(x) \right)^{1/\alpha} = \exp(-\min_{x \in \mathbb{R}^{d_1}} \mathcal{E}(x, y))$, giving the result after taking the logarithm on both sides. \square

LEMMA A.4. Define $S_{x,\delta} = \{y \in \mathbb{R}^{d_2} : \exp(\mathcal{E}(x, y)) > \exp(\max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x, y)) - \delta\}$ for any fixed $x \in \mathbb{R}^{d_1}$ and any $\delta > 0$. Let $\{\mu^\beta\}_{\beta \geq 1}$ be a family of measures in $\mathcal{P}(\mathbb{R}^{d_2})$, and assume there exists a constant $C_\delta > 0$ depending only on δ such that $\mu^\beta(S_{x,\delta}) \geq C_\delta$ for all $\beta \geq 1$. Then it holds that

$$\lim_{\beta \rightarrow \infty} \frac{1}{\beta} \log \left(\int_{\mathbb{R}^{d_2}} \exp(\beta \mathcal{E}(x, y)) d\mu^\beta(y) \right) = \max_{y \in \mathbb{R}^{d_2}} \mathcal{E}(x, y).$$

Acknowledgment. The authors sincerely thank the two referees for their careful and insightful comments and suggestions which helped improve the article.

REFERENCES

- [1] L. ARNOLD, *Stochastic Differential Equations: Theory and Applications*, Wiley-Interscience, New York, London, Sydney, 1974.
- [2] G. BORGHI, M. HERTY, AND L. PARESCHI, *A consensus-based algorithm for multi-objective optimization and its mean-field description*, in Proceedings of the 61st IEEE Conference on Decision and Control (CDC), 2022, pp. 4131–4136.
- [3] G. BORGHI, M. HERTY, AND L. PARESCHI, *Constrained consensus-based optimization*, SIAM J. Optim., 33 (2023), pp. 211–236, <https://doi.org/10.1137/22M1471304>.
- [4] S. BUBECK, *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn., 8 (2015), pp. 231–357.
- [5] J. A. CARRILLO, Y.-P. CHOI, C. TOTZECK, AND O. TSE, *An analytical framework for consensus-based global optimization method*, Math. Models Methods Appl. Sci., 28 (2018), pp. 1037–1066.

- [6] J. A. CARRILLO, S. JIN, L. LI, AND Y. ZHU, *A consensus-based global optimization method for high dimensional machine learning problems*, ESAIM Control Optim. Calc. Var., 27 (2021), S5.
- [7] J. A. CARRILLO, C. TOTZECK, AND U. VAES, *Consensus-based optimization and ensemble Kalman inversion for global optimization problems with constraints*, in Modeling and Simulation for Collective Dynamics, World Scientific, 2023, pp. 195–230.
- [8] T.-H. CHANG, M. HONG, H.-T. WAI, X. ZHANG, AND S. LU, *Distributed learning in the non-convex world: From batch data to streaming and beyond*, IEEE Signal Process. Mag., 37 (2020), pp. 26–38.
- [9] C. DASKALAKIS AND I. PANAGEAS, *The limit points of (optimistic) gradient descent in min-max optimization*, in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 9256–9266.
- [10] R. DURRETT, *Stochastic Calculus: A Practical Introduction*, Probab. Stochastics Ser., CRC Press, Boca Raton, FL, 1996.
- [11] M. FORNASIER, H. HUANG, L. PARESCHI, AND P. SÜNNEN, *Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit*, Math. Models Methods Appl. Sci., 30 (2020), pp. 2725–2751.
- [12] M. FORNASIER, H. HUANG, L. PARESCHI, AND P. SÜNNEN, *Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning*, J. Mach. Learn. Res., 22 (2021), 237.
- [13] M. FORNASIER, H. HUANG, L. PARESCHI, AND P. SÜNNEN, *Anisotropic diffusion in consensus-based optimization on the sphere*, SIAM J. Optim., 32 (2022), pp. 1984–2012, <https://doi.org/10.1137/21M140941X>.
- [14] M. FORNASIER, T. KLOCK, AND K. RIEDL, *Consensus-Based Optimization Methods Converge Globally*, preprint, arXiv:2103.15130, 2021.
- [15] M. FORNASIER, T. KLOCK, AND K. RIEDL, *Convergence of anisotropic consensus-based optimization in mean-field law*, in Applications of Evolutionary Computation, J. L. Jiménez Laredo, J. I. Hidalgo, and K. O. Babaagba, eds., Springer, Cham, 2022, pp. 738–754.
- [16] R. H. GOHARY, Y. HUANG, Z.-Q. LUO, AND J.-S. PANG, *A generalized iterative water-filling algorithm for distributed power control in the presence of a jammer*, IEEE Trans. Signal Process., 57 (2009), pp. 2660–2674.
- [17] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial networks*, Comm. ACM, 63 (2020), pp. 139–144.
- [18] E. HAZAN, *Introduction to online convex optimization*, Found. Trends Optim., 2 (2016), pp. 157–325.
- [19] D. J. HIGHAM, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Rev., 43 (2001), pp. 525–546, <https://doi.org/10.1137/S0036144500378302>.
- [20] H. HUANG AND J. QIU, *On the mean-field limit for the consensus-based optimization*, Math. Methods Appl. Sci., 45 (2022), pp. 7814–7831.
- [21] H. HUANG, J. QIU, AND K. RIEDL, *On the global convergence of particle swarm optimization methods*, Appl. Math. Optim., 88 (2023), 30.
- [22] J. KENNEDY AND R. EBERHART, *Particle swarm optimization*, in Proceedings of the ICNN’95 International Conference on Neural Networks, Vol. 4, IEEE, 1995, pp. 1942–1948.
- [23] R. A. KROHLING, F. HOFFMANN, AND L. S. COELHO, *Co-evolutionary particle swarm optimization for min-max problems using Gaussian distribution*, in Proceedings of the 2004 Congress on Evolutionary Computation, Vol. 1, IEEE, 2004, pp. 959–964.
- [24] E. C. LASKARI, K. E. PARSOPOULOS, AND M. N. VRAHATIS, *Particle swarm optimization for minimax problems*, in Proceedings of the 2002 Congress on Evolutionary Computation, Vol. 2, IEEE, 2002, pp. 1576–1581.
- [25] M. LIU, H. RAFIQUE, Q. LIN, AND T. YANG, *First-order convergence theory for weakly-convex-weakly-concave min-max problems*, J. Mach. Learn. Res., 22 (2021), 169.
- [26] Y.-F. LIU, Y.-H. DAI, AND Z.-Q. LUO, *Max-min fairness linear transceiver design for a multi-user mimo interference channel*, IEEE Trans. Signal Process., 61 (2013), pp. 2413–2423.
- [27] N. LOIZOU, H. BERARD, G. GIDEL, I. MITLIAGKAS, AND S. LACOSTE-JULIEN, *Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity*, in Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021, pp. 19095–19108.
- [28] N. LOIZOU, H. BERARD, A. JOLICOEUR-MARTINEAU, P. VINCENT, S. LACOSTE-JULIEN, AND I. MITLIAGKAS, *Stochastic Hamiltonian gradient methods for smooth games*, in Proceedings of the International Conference on Machine Learning (PMLR), 2020, pp. 6370–6381.

- [29] D. MADRAS, E. CREAGER, T. PITASSI, AND R. ZEMEL, *Learning adversarially fair and transferable representations*, in Proceedings of the International Conference on Machine Learning (PMLR), 2018, pp. 3384–3393.
- [30] A. MADRY, A. MAKELOV, L. SCHMIDT, D. TSIPRAS, AND A. VLADU, *Towards deep learning models resistant to adversarial attacks*, in Proceedings of the International Conference on Learning Representations, 2018.
- [31] E. V. MAZUMDAR, M. I. JORDAN, AND S. S. SASTRY, *On Finding Local Nash Equilibria (and Only Local Nash Equilibria) in Zero-Sum Games*, preprint, arXiv:1901.00838, 2019.
- [32] P. D. MILLER, *Applied Asymptotic Analysis*, Grad. Stud. Math. 75, American Mathematical Society, Providence, RI, 2006.
- [33] K. G. MURTY AND S. N. KABADI, *Some NP-complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.
- [34] R. B. MYERSON, *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, MA, 1997.
- [35] J. F. NASH, JR., *Equilibrium points in n -person games*, Proc. Nat. Acad. Sci. USA, 36 (1950), pp. 48–49.
- [36] M. NOUIEHED, M. SANJABI, T. HUANG, J. D. LEE, AND M. RAZAVIYAYN, *Solving a class of non-convex min-max games using iterative first order methods*, in Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 14934–14942.
- [37] B. ØKSENDAL, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer-Verlag, Berlin, 2003.
- [38] S. OMIDSHAFIEI, J. PAZIS, C. AMATO, J. P. HOW, AND J. VIAN, *Deep decentralized multi-task multi-agent reinforcement learning under partial observability*, in Proceedings of the International Conference on Machine Learning (PMLR), 2017, pp. 2681–2690.
- [39] J.-S. PANG AND G. SCUTARI, *Nonconvex games with side constraints*, SIAM J. Optim., 21 (2011), pp. 1491–1522, <https://doi.org/10.1137/100811787>.
- [40] R. PINNAU, C. TOTZECK, O. TSE, AND S. MARTIN, *A consensus-based model for global optimization and its mean-field limit*, Math. Models Methods Appl. Sci., 27 (2017), pp. 183–204.
- [41] E. PLATEN, *An introduction to numerical methods for stochastic differential equations*, in Acta Numerica, 1999, Acta Numer. 8, Cambridge University Press, Cambridge, UK, 1999, pp. 197–246.
- [42] M. RAZAVIYAYN, T. HUANG, S. LU, M. NOUIEHED, M. SANJABI, AND M. HONG, *Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances*, IEEE Signal Process. Mag., 37 (2020), pp. 55–66.
- [43] K. RIEDL, *Leveraging memory effects and gradient information in consensus-based optimization: On global convergence in mean-field law*, European J. Appl. Math., (2023), pp. 1–32, <https://doi.org/10.1017/S0956792523000293>.
- [44] Y. SHI AND R. A. KROHLING, *Co-evolutionary particle swarm optimization to solve min-max problems*, in Proceedings of the 2002 Congress on Evolutionary Computation, Vol. 2, IEEE, 2002, pp. 1682–1687.
- [45] M. SION, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.
- [46] J. VON NEUMANN, *Zur Theorie der Gesellschaftsspiele*, Math. Ann., 100 (1928), pp. 295–320.
- [47] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 2007.

License for [CBO-SP].

The permission to reprint and include the material is printed on the next page(s).

Von: Kelly Thomas Thomas@siam.org
Betreff: RE: Request for Permission to use Material in my Dissertation (SIOPT M152780)
Datum: 14. März 2024 um 21:14
An: Konstantin Riedl konstantin.riedl@ma.tum.de

Dear Mr. Riedl:

SIAM is happy to give permission to reprint material from “Consensus-Based Optimization Methods Converge Globally“ (SIOPT M152780) and “Consensus-Based Optimization for Saddle Point Problems“ (SICON M154336) in your dissertation. Please acknowledge the original publications, using the complete bibliographic information if available.

Sincerely,

Kelly Thomas
Managing Editor
Society for Industrial and Applied Mathematics
3600 Market Street - 6th Floor
Philadelphia, PA 19104
thomas@siam.org / (267) 350-6387

Von: Konstantin Riedl konstantin.riedl@ma.tum.de
Betreff: Request for Permission to use Material in my Dissertation (SICON M154336)
Datum: 14. März 2024 um 18:00
An: Thomas@siam.org

Dear Mrs. Thomas,

As one of the authors of the article „Consensus-Based Optimization for Saddle Point Problems“ (SICON M154336) which is accepted for publication in the *SIAM Journal on Control and Optimization*, I am reaching out to you as the Managing Editor of SIAM to request permission to include the paper in my dissertation (doctoral thesis). Since I am pursuing a cumulative dissertation, it is necessary by the rules of my university, the Technical University of Munich, and the School of Computation, Information and Technology, that I provide and include in my dissertation a **written letter of approval from the publisher** for all my publications that are part of my dissertation.

I hereby kindly ask for such a confirmation from SIAM (by email or a weblink to your terms and conditions), which allows me to use the aforementioned article in my dissertation.

If you have any questions beforehand, do not hesitate to ask.

Best regards,
Konstantin

Konstantin Riedl, M.Sc.

Technical University of Munich
School of Computation, Information and Technology
Department of Mathematics
Chair for Applied Numerical Analysis

Munich Center for Machine Learning

Institute for Ethics in Artificial Intelligence

Email: konstantin.riedl@ma.tum.de



It has been a journey.
It was supposed to be a journey.

Konstantin Riedl
Munich, 2024