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 metaheuristic derivative-free optimization method that can globally minimize nonconvex nonsmooth 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 $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ and the search for an associated globally

*Email: massimo.fornasier@ma.tum.de
†Email: tmklock@googlemail.com
‡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, 5, 7, 44, 57]. 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 [56], Evolutionary Programming [26], the Metropolis-Hastings algorithm [35], Genetic Algorithms [37], Particle Swarm Optimization [44], 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., [38, 65]. While transferring some of the ideas of [65] 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 [68].

Recently, the authors of [12, 54] 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 $\widehat{\rho}_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{V_t^i}$ converges to a Dirac delta $\delta_{\widetilde{v}}$ at some $\widetilde{v} \in \mathbb{R}^d$ close to v^*.

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

$$\begin{aligned} V^i_{(k+1)\Delta t} - V^i_{k\Delta t} = &- \Delta t \lambda \left(V^i_{k\Delta t} - v_\alpha(\widehat{\rho}^N_{k\Delta t})\right) H\left(\mathcal{E}(V^i_{k\Delta t}) - \mathcal{E}\left(v_\alpha(\widehat{\rho}^N_{k\Delta t})\right)\right) \\ &+ \sigma \left\|V^i_{k\Delta t} - v_\alpha(\widehat{\rho}^N_{k\Delta t})\right\|_2 B^i_{k\Delta t}, \end{aligned} \tag{1}$$

$$V^i_0 \sim \rho_0 \quad \text{for all } i = 1, \dots, N, \tag{2}$$

where $((B^i_{k\Delta t})_{k=0,\dots,K-1})_{i=1,\dots,N}$ are independent, identically distributed Gaussian random vectors in $\mathbb{R}^d$ with zero mean and covariance matrix $\Delta t \mathsf{Id}_d$. 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 (1) originates from a simple Euler-Maruyama time discretization [36, 55] of the system of stochastic differential equations (SDEs)

$$dV_t^i = -\lambda \left(V_t^i - v_\alpha(\widehat{\rho}_t^N)\right) H\left(\mathcal{E}(V_t^i) - \mathcal{E}\left(v_\alpha(\widehat{\rho}_t^N)\right)\right) dt + \sigma \left\|V_t^i - v_\alpha(\widehat{\rho}_t^N)\right\|_2 dB_t^i, \tag{3}$$

$$V_0^i \sim \rho_0 \quad \text{for all } i = 1, \dots, N, \tag{4}$$

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 (1) and (3) 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

$$v_\alpha(\widehat{\rho}_t^N) := \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\widehat{\rho}_t^N)}} d\widehat{\rho}_t^N(v), \quad \text{with} \quad \omega_\alpha(v) := \exp(-\alpha \mathcal{E}(v)). \tag{5}$$

Definition (5) is motivated by the well-known Laplace principle [23, 51, 54], which states that, for any absolutely continuous probability distribution ϱ on $\mathbb{R}^d$, we have

$$\lim_{\alpha\to\infty} \left(-\frac{1}{\alpha} \log \left(\int \omega_\alpha(v)\, d\varrho(v) \right) \right) = \inf_{v \in \mathrm{supp}(\varrho)} \mathcal{E}(v). \tag{6}$$

Alternatively, we can also interpret (5) as an approximation of $\arg\min_{i=1,\dots,N} \mathcal{E}(V_t^i)$, which improves as $\alpha \to \infty$, provided the minimizer uniquely exists. The univariate function $H : \mathbb{R} \to [0, 1]$ appearing in the first term of (1) and (3) 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 (1) and (3) 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 [10, 12–14, 18, 27–30, 42, 45, 67], 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 [13, 28]. Moreover, several variants of the dynamics have been proposed, such as ones integrating memory mechanisms [59, 67] or others using jump-diffusion processes [42]. 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 [13, 30] 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 [13, 30, 59] where the method is used for training a shallow and a convolutional neural network for image classification of the MNIST database of handwritten digits [46], to the recent paper [15] where CBO is used in the setting of clustered federated learning, to [59] where a compressed sensing problem is solved, or to the line of works [27–29] where (1) and (3) 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 [10, 14].

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 $\widehat{\rho}_t^N \to \delta_{\widetilde{v}}$ occurs, and on quantitatively bounding the associated errors $\mathcal{E}(\widetilde{v}) - \underline{\mathcal{E}}$ and $\|\widetilde{v} - v^*\|_2$. A theoretical analysis of the dynamics can either be done on the microscopic systems (1) or (3), as for instance in [33, 34], or, as in [12, 54], by analyzing the macroscopic behavior of the agent density through a mean-field limit associated with the particle-based dynamics (3), given, for initial data $\overline{V}_0 \sim \rho_0$, by

$$d\overline{V}_t = -\lambda \left(\overline{V}_t - v_\alpha(\rho_t)\right) H\big(\mathcal{E}(\overline{V}_t) - \mathcal{E}(v_\alpha(\rho_t))\big)\, dt + \sigma \left\|\overline{V}_t - v_\alpha(\rho_t)\right\|_2 dB_t, \tag{7}$$

where $\rho_t = \mathrm{Law}(\overline{V}_t)$. The weak convergence of the microscopic system (3) to the mean-field limit (7), or, more precisely, of the empirical measure $\widehat{\rho}_t^N$ to ρ_t as $N \to \infty$, has been shown recently in [39], see also Remark 1.2 for additional details. This legitimates to analyze (7) in lieu of (3). The measure $\rho \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^d))$ with $\rho_t = \rho(t) = \mathrm{Law}(\overline{V}_t)$ satisfies the nonlinear nonlocal Fokker-Planck equation

$$\partial_t \rho_t = \lambda \mathrm{div}\big((v - v_\alpha(\rho_t))\, H(\mathcal{E}(v) - \mathcal{E}(v_\alpha(\rho_t)))\, \rho_t\big) + \frac{\sigma^2}{2} \Delta\big(\|v - v_\alpha(\rho_t)\|_2^2\, \rho_t\big) \tag{8}$$

in a weak sense (see Definition 3.1). Leveraging this partial differential equation (PDE), the authors of [12, 54] analyze the large time behavior of the particle density $t \mapsto \rho_t$ instead of the microscopic systems (1) and (3). Studying the mean-field limit (7) or (8) 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 (3) to the mean-field limit (7) as the number of agents grows. We discuss results about the mean-field approximation in Remark 1.2 and make it rigorous in Proposition 3.11. Hence, in view of its validity and as already done in the preceding works [12, 54], in the first part of the paper we concentrate on establishing convergence in mean-field law for (3), as defined in Definition 1.1 below. That is, we analyze the mean-field dynamics (7) and (8) in place of the interacting particle system (3). 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 (1).

Definition 1.1 (Convergence in mean-field law)**.** *Let $F, G : \mathcal{P}(\mathbb{R}^d) \otimes \mathbb{R}^d \to \mathbb{R}^d$ be two functions and consider for $i = 1, \dots, N$ the SDEs expressed in Itô's form as*

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

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

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

satisfy $\lim_{t\to\infty} W_p\left(\rho_t, \delta_{\widetilde{v}}\right) = 0$ 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 1.2 (Mean-field approximation)**.** The definition of convergence in mean-field law as given in Definition 1.1 is justified as follows: As the number of agents N in the interacting particle system (3) 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 (7). This phenomenon is known as the mean-field approximation. More formally, as $N \to \infty$, we expect the empirical measure $\widehat{\rho}_t^N$ to converge in law to ρ_t for almost every t, see [41, Definition 1]. The classical way to establish such mean-field approximation is to prove, by means of the coupling method, propagation of chaos [49, 66], as implied for instance by

$$\max_{i=1,\dots,N} \sup_{t\in[0,T]} \mathbb{E}\big\|V_t^i - \overline{V}_t^i\big\|_2^2 \leq CN^{-1}, \tag{9}$$

where $\overline{V}^i$ denote N i.i.d. copies of the mean-field dynamics (7), 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 [16, 17]. 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 (9) has been left as a difficult and open problem in [12, 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 [17, 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 [39] show that the empirical random particle measure $\widehat{\rho}^N$ associated with the dynamics (3) converges in distribution to the deterministic particle distribution $\rho \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^d))$ satisfying (8). 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 (9), it closes the mean-field limit gap qualitatively. A desired quantitative result has been established recently in [27, Theorem 3.1 and Remark 3.1] for a variant of the microscopic system (3) supported on a compact hypersurface Γ. In [27] the weak convergence of the variant of (3) 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

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

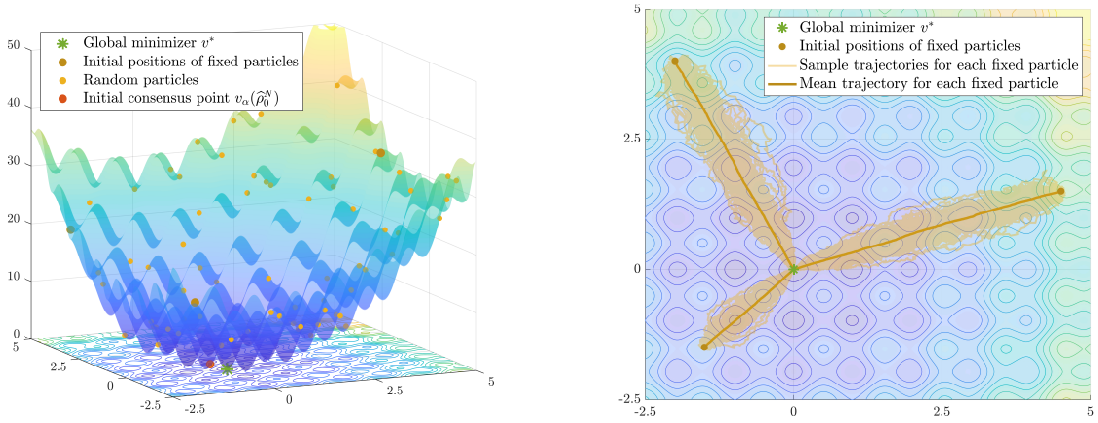
The obtained convergence rate reads CN^{-1} with C depending in particular on

$$C_\alpha := \exp\big(\alpha\big(\sup\nolimits_{v\in\Gamma} \mathcal{E}(v) - \inf\nolimits_{v\in\Gamma} \mathcal{E}(v)\big)\big).$$

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 (9). Leveraging the techniques from [27] and the boundedness of moments established in [12, Lemma 3.4], we provide in Proposition 3.11 below a result of the form (9) on the plane $\mathbb{R}^d$ which holds with high probability. A more refined analysis conducted recently by the authors of [31], 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 (9) 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 [31, Theorem 2.6].

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 (7) and (8). Nevertheless, as a consequence thereof, we return to the analysis of the numerical scheme (1) and its global convergence in Section 3.3.

Contributions 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



(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 (1)–(2), with parameters $\Delta t = 0.01$, $\alpha = 10^{15}$, $\lambda = 1$ and $\sigma = 0.1$, and $N = 32000$ agents initialized according to $\rho_0 = \mathcal{N}\big((8,8),20\big)$. 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.

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 [12, 33, 34] 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 $N \to \infty$. Secondly, in consideration of its central significance with regards to the computational complexity of the numerical scheme (1) we establish a novel probabilistic quantitative result about the convergence of the interacting particle system (3) to the corresponding mean-field limit (8), 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 (1) 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., [10, 13, 28–30, 42], 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., [22, 32, 40]. While the present paper has foundational and theoretical nature and aims at completely clarifying the convergence of the numerical scheme (1) 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

[13, 15, 18, 28–30, 59].

Remark 1.3. 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., [20, 24, 48]. Recent work [60, 61] 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 [19, 20, 25, 62] or Stein Variational Gradient Descent [47], 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-Łojasiewicz-like conditions [43] or certain families of log-Sobolev inequalities [20] 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 [20] 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 [21, 50, 63, 64] 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 $\big(\mathcal{P}_2(\mathbb{R}^d), W_2\big)$, 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 [21, 50, 63, 64] 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 (8) describing the mean-field behavior of CBO cannot be understood as a gradient flow of any energy on $\big(\mathcal{P}_2(\mathbb{R}^d), W_2\big)$. 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 (7), respectively, the Fokker-Planck equation (8), before presenting and discussing the main result about the convergence of the dynamics (7) and (8) 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 (1), 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.

For the sake of reproducible research, in the GitHub repository https://github.com/KonstantinRiedl/CBOGlobalConvergenceAnalysis we provide the Matlab code implementing the CBO algorithm analyzed in this work and used to create all visualizations. Python and Julia code for CBO [6] can be also found in the GitHub repositories https://github.com/PdIPS/CBXpy and https://github.com/PdIPS/ConsensusBasedX.jl.

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 $f : X \to Y$ 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 [12], and related works [33, 34], which prove convergence as a consequence of a monotonous decay of the variance of ρ_t and by employing the asymptotic Laplace principle (6). 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 $\mathcal{E}(\widetilde{v}) - \mathcal{E}(v^*)$ 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 [12], which has been adapted in other works, e.g., [28, 33, 34], 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 $\delta_{\widetilde{v}}$ for some $\widetilde{v}$ close to v^*. The analysis in [12] proves this under certain assumptions by first showing that ρ_t converges to a Dirac delta around *some* $\widetilde{v} \in \mathbb{R}^d$ and then concluding $\widetilde{v} \approx v^*$ in a subsequent step. Regarding the first step, the authors of [12] study the variance of ρ_t, defined as $\mathrm{Var}(\rho_t) := \frac{1}{2}\int \|v - \mathbb{E}(\rho_t)\|_2^2 \, d\rho_t(v)$, where $\mathbb{E}(\rho_t) := \int v \, d\rho_t(v)$, and show that $\mathrm{Var}(\rho_t)$ decays exponentially fast in t under a well-preparedness assumption about the initial condition ρ_0. More precisely, in [12, Section 4.1] the authors use Itô's lemma to derive for the time-evolution of $\mathrm{Var}(\rho_t)$ the expression

$$\frac{d}{dt}\mathrm{Var}(\rho_t) = -\left(2\lambda - d\sigma^2\right)\mathrm{Var}(\rho_t) + \frac{d\sigma^2}{2}\|\mathbb{E}(\rho_t) - v_\alpha(\rho_t)\|_2^2 . \tag{10}$$

For parameter choices $2\lambda > d\sigma^2$, the first term in (10) is negative and one could *almost* apply Grönwall's inequality to obtain the asserted exponential decay of $\mathrm{Var}(\rho_t)$. 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 $\alpha \to 0$ the weight function $\omega_\alpha(v) = \exp(-\alpha\mathcal{E}(v))$

associated with $v_\alpha(\rho_t)$ converges to 1 pointwise and consequently $v_\alpha(\rho_t) \to \mathbb{E}(\rho_t)$. 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 [12] compromise to assume that the initial distribution ρ_0 satisfies the well-preparedness assumptions

$$\alpha e^{-2\alpha\underline{\mathcal{E}}}(\sigma^2+2\lambda) < 3/8 \quad \text{and} \quad 2\lambda\|\omega_\alpha\|^2_{L_1(\rho_0)} - \mathrm{Var}\,(\rho_0) - 2d\sigma^2\,\|\omega_\alpha\|_{L_1(\rho_0)}\, e^{-\alpha\underline{\mathcal{E}}} \geq 0. \tag{11}$$

Since ρ_t evolves from ρ_0 according to the Fokker-Planck equation (8), these conditions restrict ρ_t and allow for bounding $\|\mathbb{E}(\rho_t) - v_\alpha(\rho_t)\|_2$ by a suitable multiple of $\mathrm{Var}\,(\rho_t)$. The exponential decay of $\mathrm{Var}\,(\rho_t)$ then follows from (10) after applying Grönwall's inequality, see [12, Theorem 4.1]. Furthermore, the conditions in (11) also allow for proving convergence of ρ_t to a stationary Dirac delta at $\widetilde{v} \in \mathbb{R}^d$.

Given convergence to a Dirac at $\widetilde{v}$, in a second step it is shown $\mathcal{E}(\widetilde{v}) \approx \mathcal{E}(v^*)$. 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 $-\frac{1}{\alpha}\log\big(\|\omega_\alpha\|_{L_1(\rho_t)}\big) \leq -\frac{1}{\alpha}\log\big(\|\omega_\alpha\|_{L_1(\rho_0)}\big) + \frac{\varepsilon}{2}$. 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 [12, Lemma 4.1], which is then combined with the formerly proven exponentially decaying variance, see [12, Proof of Theorem 4.2]. Then, by recognizing that the Laplace principle (6) implies the existence of some $\alpha \gg 1$ with

$$-\frac{1}{\alpha}\log\big(\|\omega_\alpha\|_{L_1(\rho_0)}\big) - \underline{\mathcal{E}} < \frac{\varepsilon}{2}, \tag{12}$$

and by establishing the convergence $\|\omega_\alpha\|_{L_1(\rho_t)} \to \exp(-\alpha\mathcal{E}(\widetilde{v}))$ as $t \to \infty$, one obtains the desired result $\mathcal{E}(\widetilde{v}) - \underline{\mathcal{E}} < \varepsilon$ in the limit $t \to \infty$, see [12, Lemma 4.2]. The gap $\mathcal{E}(\widetilde{v}) - \underline{\mathcal{E}}$ 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 $\mathcal{E}(\widetilde{v}) - \underline{\mathcal{E}} < \varepsilon$. In fact, $\alpha(\varepsilon)$ may grow uncontrollably as we decrease the accuracy ε. The first step, however, requires the conditions in (11) which, in the most optimistic case, where $\sigma = 0$, imply

$$\mathrm{Var}\,(\rho_0) \leq \frac{3}{8\alpha}\left(\int \exp\big(-\alpha(\mathcal{E}(v) - \underline{\mathcal{E}})\big)\, d\rho_0(v)\right)^2. \tag{13}$$

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 (13) to hold for generic choices such as a uniform distribution on a compact set.

We add that the works [33, 34] conduct a similarly flavored analysis for the fully time-discretized microscopic system (1), 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., $(B^i_{k\Delta t})_{k=1,\dots,K} = (B_{k\Delta t})_{k=1,\dots,K}$ for all $i = 1, \dots, N$. Such a choice leads to a less explorative dynamics, but it simplifies the consensus formation analysis. For proving $\mathcal{E}(\widetilde{v}) \approx \underline{\mathcal{E}}$ however, the authors again require an initial configuration ρ_0 that satisfies a technical concentration condition like (12), see for instance [34, 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 (7), in the case $H \equiv 1$, becomes

$$\frac{d}{dt}\mathbb{E}\big[\overline{V}_t\big|\overline{V}_0\big] = -\lambda\mathbb{E}\big[\big(\overline{V}_t - v_\alpha(\rho_t)\big)\big|\overline{V}_0\big] = -\lambda\mathbb{E}\big[\big(\overline{V}_t - v^*\big)\big|\overline{V}_0\big] + \lambda\left(v_\alpha(\rho_t) - v^*\right). \tag{14}$$

Furthermore, if $\mathcal{E}$ is locally Lipschitz continuous and satisfies the coercivity condition

$$\|v - v^*\|_2 \leq \frac{1}{\eta}\big(\mathcal{E}(v) - \mathcal{E}(v^*)\big)^\nu = \frac{1}{\eta}\big(\mathcal{E}(v) - \underline{\mathcal{E}}\big)^\nu, \quad \text{for all } v \in \mathbb{R}^d, \tag{15}$$

and for some $\eta > 0$ and $\nu \in (0, \infty)$, the second term on the right-hand side of (14) 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 $\overline{V}_t$ is well-approximated by

$$\frac{d}{dt}\mathbb{E}\left[\overline{V}_t|\overline{V}_0\right] \approx -\lambda\mathbb{E}\left[\big(\overline{V}_t - v^*\big)|\overline{V}_0\right], \tag{16}$$

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 $\mathcal{V} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$, given by

$$\mathcal{V}(\rho_t) := \frac{1}{2}\int \|v - v^*\|_2^2 \, d\rho_t(v). \tag{17}$$

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 $\mathcal{V}(\rho_t) \to 0$ 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 4.1 for details), $\mathcal{V}(\rho_t)$ follows an evolution similar to (10), with $\mathrm{Var}\,(\rho_t)$ 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 $\mathrm{Var}\,(\rho_t)$ 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 $\|v_\alpha(\rho_t) - v^*\|_2 \to 0$ under (15) as $\alpha \to \infty$ and we can even establish (see Proposition 4.5 for details) the quantitative estimate

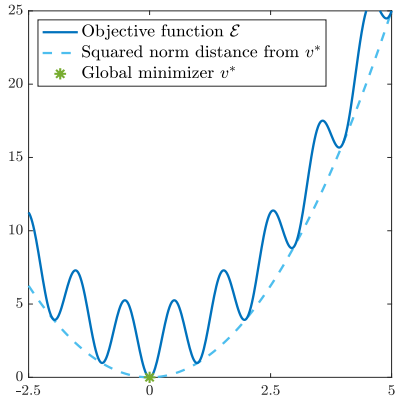
$$\|v_\alpha(\varrho) - v^*\|_2 \leq \frac{(2Lr)^\nu}{\eta} + \frac{\exp\left(-\alpha L r\right)}{\varrho(B_r(v^*))}\int \|v - v^*\|_2 \, d\varrho(v)$$

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.

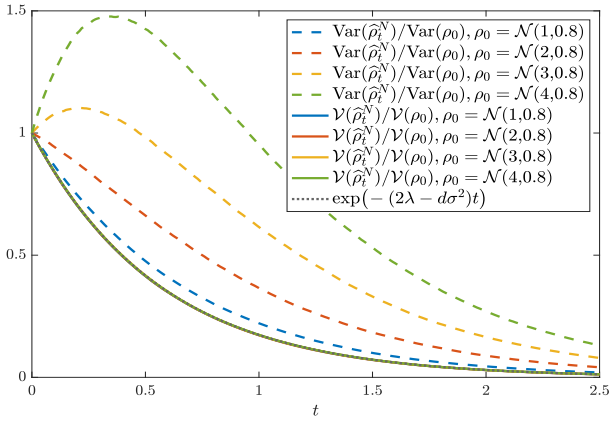
Finally, we note that $\mathcal{V}(\rho_t)$ majorizes $\mathrm{Var}\,(\rho_t)$ because $u \mapsto \frac{1}{2}\int \|v - u\|_2^2 \, d\rho_t(v)$ is minimized by the expectation $\mathbb{E}(\rho_t)$. This relation may be a source of concern, as it shows that proving $\mathcal{V}(\rho_t) \to 0$ implies $\mathrm{Var}\,(\rho_t) \to 0$. We emphasize however that this does not imply a majorization for the corresponding time derivatives. In fact, Example 2.1 suggests that $\mathcal{V}(\rho_t)$ can decay exponentially while $\mathrm{Var}\,(\rho_t)$ increases initially.

Example 2.1. 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 (1) using $N = 320000$

agents, discrete time step size $\Delta t = 0.01$ and parameters $\alpha = 10^{15}$ (i.e., the consensus point is the arg min 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 $\mathrm{Var}(\widehat{\rho}_t^N)$ (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 (13), 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 3.7.



(a) The Rastrigin function and the map $v \mapsto \|v - v^*\|_2^2$



(b) Evolution of the variance $\mathrm{Var}(\widehat{\rho}_t^N)$ and $\mathcal{V}(\widehat{\rho}_t^N)$ for different initial conditions ρ_0

Figure 2: **(a)** The Rastrigin function as objective function $\mathcal{E}$ and the squared Euclidean distance from v^*. **(b)** The evolution of the variance $\mathrm{Var}(\widehat{\rho}_t^N)$ 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 (1) 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, $\mathrm{Var}(\widehat{\rho}_t^N)$ 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 (7), respectively, the associated Fokker-Planck equation (8). 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 (7) and (8) 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 (1) 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 (8).

Definition 3.1. *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 (8) 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* $t \in (0,T)$

$$\begin{aligned}\frac{d}{dt}\int \phi(v)\,d\rho_t(v) = &-\lambda \int H(\mathcal{E}(v)-\mathcal{E}(v_\alpha(\rho_t)))\,\langle v - v_\alpha(\rho_t), \nabla\phi(v)\rangle\,d\rho_t(v)\\ &+\frac{\sigma^2}{2}\int \|v-v_\alpha(\rho_t)\|_2^2\,\Delta\phi(v)\,d\rho_t(v)\end{aligned} \tag{18}$$

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

If the cutoff function H in the dynamics (7) is inactive, i.e., satisfies $H \equiv 1$, the authors of [12] prove the following well-posedness result.

Theorem 3.2 ([12, Theorems 3.1, 3.2])**.** *Let* $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. *Let* $H \equiv 1$ *and consider* $\mathcal{E}: \mathbb{R}^d \to \mathbb{R}$ *with* $\underline{\mathcal{E}} > -\infty$, *which, for constants* $C_1, C_2 > 0$, *satisfies*

$$|\mathcal{E}(v)-\mathcal{E}(w)| \le C_1(\|v\|_2 + \|w\|_2)\,\|v-w\|_2\,, \quad \text{for all } v,w \in \mathbb{R}^d, \tag{19}$$

$$\mathcal{E}(v) - \underline{\mathcal{E}} \le C_2(1+\|v\|_2^2), \quad \text{for all } v \in \mathbb{R}^d. \tag{20}$$

If in addition, either $\sup_{v\in\mathbb{R}^d}\mathcal{E}(v) < \infty$, *or* $\mathcal{E}$ *satisfies for some constants* $C_3, C_4 > 0$

$$\mathcal{E}(v) - \underline{\mathcal{E}} \ge C_3\,\|v\|_2^2\,, \quad \text{for all } \|v\|_2 \ge C_4, \tag{21}$$

then there exists a unique nonlinear process $\overline{V} \in \mathcal{C}([0,T],\mathbb{R}^d)$ *satisfying* (7) *in the strong sense. The associated law* $\rho = \mathrm{Law}(\overline{V})$ *has regularity* $\rho \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^d))$ *and is a weak solution to the Fokker-Planck equation* (8).

Remark 3.3. The regularity $\rho \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^d))$ stated in Theorem 3.2, and also obtained in Theorem 3.4 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 [12, 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 3.1. Namely, if $\rho \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^d))$ solves (8) in the weak sense, Identity (18) 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) $\|\nabla\phi(v)\|_2 \le C(1+\|v\|_2)$ 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 3.2, we can extend the existence of solutions to an active Lipschitz-continuous cutoff function H.

Theorem 3.4. *Let* $H \not\equiv 1$ *be* L_H*-Lipschitz continuous. Then, under the assumptions of Theorem 3.2, there exists a nonlinear process* $\overline{V} \in \mathcal{C}([0,T],\mathbb{R}^d)$ *satisfying* (7) *in the strong sense. The associated law* $\rho = \mathrm{Law}(\overline{V})$ *has regularity* $\rho \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^d))$ *and is a weak solution to the Fokker-Planck equation* (8).

Proof sketch. 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 [4, Chapter 6], we can uniquely solve the auxiliary SDE

$$d\widetilde{V}_t = -\lambda\big(\widetilde{V}_t - u_t\big)H\big(\mathcal{E}(\widetilde{V}_t) - \mathcal{E}(u_t)\big)\,dt + \sigma\big\|\widetilde{V}_t - u_t\big\|_2\,dB_t \quad \text{with} \quad \widetilde{V}_0 \sim \rho_0,$$

as the coefficients are locally Lipschitz continuous and have at most linear growth, due to the assumptions on $\mathcal{E}$ and H. This induces $\widetilde{\rho}_t = \mathrm{Law}(\widetilde{V}_t)$. Moreover, the regularity of the initial distribution $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ allows for a fourth-order moment estimate of the form $\mathbb{E}\|\widetilde{V}_t\|_2^4 \le \big(1+\mathbb{E}\|\widetilde{V}_0\|_2^4\big)e^{ct}$, see, e.g. [4, Chapter 7]. So, in particular, $\widetilde{\rho} \in \mathcal{C}([0,T],\mathcal{P}_4(\mathbb{R}^d))$.
Step 2: Let us now define, for some constant $C > 0$, the test function space

$$\mathcal{C}_*^2(\mathbb{R}^d) := \big\{\phi \in \mathcal{C}^2(\mathbb{R}^d) : \|\nabla\phi(v)\|_2 \le C(1+\|v\|_2) \text{ and } \sup_{v\in\mathbb{R}^d}|\Delta\phi(v)| < \infty\big\}. \tag{22}$$

For some $\phi \in \mathcal{C}^2_*(\mathbb{R}^d)$, by Itô's formula, we derive

$$\begin{aligned} d\phi(\widetilde{V}_t) = \nabla\phi(\widetilde{V}_t) \cdot \big(-\lambda\big(\widetilde{V}_t - u_t\big) H\big(\mathcal{E}(\widetilde{V}_t) - \mathcal{E}(u_t)\big)\, dt + \sigma \big\|\widetilde{V}_t - u_t\big\|_2\, dB_t\big) \\ + \frac{1}{2}\sigma^2 \Delta\phi(\widetilde{V}_t) \big\|\widetilde{V}_t - u_t\big\|_2^2\, dt. \end{aligned}$$

After taking the expectation, applying Fubini's theorem and observing that the stochastic integral $\mathbb{E}\int_0^t \nabla\phi(\widetilde{V}_t) \cdot \|\widetilde{V}_t - u_t\|_2\, dB_t$ vanishes as a consequence of [53, Theorem 3.2.1(iii)] due to the established regularity $\widetilde{\rho} \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^d))$ and $\phi \in \mathcal{C}^2_*(\mathbb{R}^d)$, we obtain

$$\frac{d}{dt}\mathbb{E}\phi(\widetilde{V}_t) = -\lambda \mathbb{E}\nabla\phi(\widetilde{V}_t) \cdot \big(\widetilde{V}_t - u_t\big) H\big(\mathcal{E}(\widetilde{V}_t) - \mathcal{E}(u_t)\big) + \frac{\sigma^2}{2}\mathbb{E}\Delta\phi(\widetilde{V}_t)\big\|\widetilde{V}_\tau - u_t\big\|_2^2$$

according to the fundamental theorem of calculus. This shows that the measure $\widetilde{\rho} \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^d))$ satisfies the Fokker-Planck equation

$$\begin{aligned} \frac{d}{dt}\int \phi(v)\, d\widetilde{\rho}_t(v) = -\lambda \int H(\mathcal{E}(v) - \mathcal{E}(u_t))\, \langle v - u_t, \nabla\phi(v)\rangle\, d\widetilde{\rho}_t(v) \\ + \frac{\sigma^2}{2}\int \|v - u_t\|_2^2\, \Delta\phi(v)\, d\widetilde{\rho}_t(v). \end{aligned} \tag{23}$$

The remainder is identical to the cited paper and is summarized briefly for completeness.
Step 3: Setting $\mathcal{T}u := v_\alpha(\widetilde{\rho}) \in \mathcal{C}([0,T], \mathbb{R}^d)$ provides the self-mapping property of the map

$$\mathcal{T} : \mathcal{C}([0,T], \mathbb{R}^d) \to \mathcal{C}([0,T], \mathbb{R}^d), \quad u \mapsto \mathcal{T}u = v_\alpha(\widetilde{\rho}),$$

which is compact as a consequence of both a stability estimate for the consensus point showing that $\|v_\alpha(\widetilde{\rho}_t) - v_\alpha(\widetilde{\rho}_s)\|_2 \lesssim W_2(\widetilde{\rho}_t, \widetilde{\rho}_s)$ for $\widetilde{\rho}_t, \widetilde{\rho}_s \in \mathcal{P}_4(\mathbb{R}^d)$ [12, Lemma 3.2] and the Hölder-1/2 continuity of the Wasserstein-2 distance $W_2(\widetilde{\rho}_t, \widetilde{\rho}_s)$.
Step 4: Finally, for $u = \vartheta\mathcal{T}u$ with $\vartheta \in [0,1]$, there exists $\rho \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^d))$ satisfying (23) such that $u = \vartheta v_\alpha(\rho)$, for which a uniform bound can be obtained either due to the boundedness or the growth condition of $\mathcal{E}$. An application of the Leray-Schauder fixed point theorem concludes the proof by providing a solution to (7). □

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 3.5 (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*

$$\|v - v^*\|_2 \le (\mathcal{E}(v) - \underline{\mathcal{E}})^\nu/\eta \quad \text{for all } v \in B_{R_0}(v^*), \tag{24}$$

$$\mathcal{E}(v) - \underline{\mathcal{E}} > \mathcal{E}_\infty \quad \text{for all } v \in \big(B_{R_0}(v^*)\big)^c. \tag{25}$$

Furthermore, for the case $H \not\equiv 1$, we additionally require that $\mathcal{E}$ fulfills a local Lipschitz continuity-like condition, i.e.,

A3 *there exist $L_\mathcal{E} > 0$ and $\gamma \ge 0$ such that*

$$\mathcal{E}(v) - \underline{\mathcal{E}} \le L_\mathcal{E}(1 + \|v - v^*\|_2^\gamma)\, \|v - v^*\|_2 \quad \text{for all } v \in \mathbb{R}^d. \tag{26}$$

Remark 3.6. The analyses in [12] 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 3.5 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 (24), 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 [28], as a quadratic growth condition in the case $\nu = 1/2$ from [3, 52], or as the Hölderian error bound condition in the case $\nu \in (0, 1]$ [8]. In [52, Theorem 4] and [43, Theorem 2] many equivalent or stronger conditions are identified to imply Equation (24) globally on $\mathbb{R}^d$. Furthermore, in [28, 69], (24) is shown to hold globally for objectives related to various machine learning problems. The second part of A2, Equation (25), 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 $\mathcal{E}_\infty$ as $\|v\|_2 \to \infty$, because such functions do not satisfy the growth condition (24) globally. However, whenever (24) holds globally, we take $R_0 = \infty$, i.e., $B_{R_0}(v^*) = \mathbb{R}^d$ and (25) is void. We also note that (24) and (25) imply the uniqueness of the global minimizer v^* on $\mathbb{R}^d$.

Finally, to cover the active cutoff case $H \not\equiv 1$, 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.

We are now ready to state the main result. The proof is deferred to Section 4.

Theorem 3.7. *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 $v^* \in \operatorname{supp}(\rho_0)$. Define $\mathcal{V}(\rho_t)$ as given in (17). 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*

$$T^* := \frac{1}{(1-\vartheta)\left(2\lambda - d\sigma^2\right)} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right). \tag{27}$$

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 (8) on the time interval $[0, T^*]$ with initial condition ρ_0, we have

$$\mathcal{V}(\rho_T) = \varepsilon \quad \text{with} \quad T \in \left[\frac{1-\vartheta}{(1+\vartheta/2)} T^*, T^*\right]. \tag{28}$$

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

$$W_2^2(\rho_t, \delta_{v^*}) = 2\mathcal{V}(\rho_t) \leq 2\mathcal{V}(\rho_0) \exp\left(-(1-\vartheta)\left(2\lambda - d\sigma^2\right) t\right). \tag{29}$$

If $\mathcal{E}$ additionally satisfies A3, the same conclusion holds for any $H : \mathbb{R}^d \to [0, 1]$ that satisfies $H(x) = 1$ whenever $x \geq 0$.

The assumption $v^* \in \operatorname{supp}(\rho_0)$ 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 (8), see Remark 4.7. 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 (29) 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 $\alpha \to \infty$ to allow for $\vartheta \to 0$. 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 [13, 30].

3.3 Global convergence in probability

To stress the relevance of the main result of this paper, Theorem 3.7, we now show how Estimate (29) plays a fundamental role in establishing a quantitative convergence result for the numerical scheme (1) to the global minimizer v^*. By paying the price of having a probabilistic statement about the convergence of CBO as in Theorem 3.8, 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 3.8. *Fix $\varepsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Then, under the assumptions of Theorem 3.7 and Proposition 3.11, and with $K := T/\Delta t$, where T is as in (28), the iterations $((V^i_{k\Delta t})_{k=0,\dots,K})_{i=1,\dots,N}$ generated by the numerical scheme (1) converge in probability to v^*. More precisely, the empirical mean of the final iterations fulfills*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} V^i_{K\Delta t} - v^* \right\|_2^2 \leq \varepsilon_{\text{total}} \tag{30}$$

with probability larger than $1 - \big(\delta + \varepsilon_{\text{total}}^{-1}(6C_{\text{NA}}(\Delta t)^{2m} + 3C_{\text{MFA}}N^{-1} + 12\varepsilon)\big)$. 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 3.7. 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 (1) together with some implementational aspects which allow to reduce the overall runtime of the algorithm in practice.

Remark 3.9 (Computational complexity)**.** To achieve Estimate (30) with probability of at least $(1-2\delta)$, the implementable CBO scheme (1) 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

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

iterations. Here, the parameter dependence of C_{NA} and C_{MFA} is as described in Theorem 3.8. 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 (1) 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 [13, 30] 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 [28] 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 3.8, 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

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

Throughout this section, $M > 0$ denotes a constant which we shall adjust at the end of the proof of Theorem 3.8. Before stating the mean-field approximation result, Proposition 3.11, let us estimate the measure of the set Ω_M in Lemma 3.10. The proofs of both statements are deferred to Section 5.

Lemma 3.10. *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 (3) and let $((\overline{V}_t^i)_{t\geq 0})_{i=1,\dots,N}$ be N independent copies of the strong solution to the mean-field dynamics (7). Then, under the assumptions of Theorem 3.2, for any $M > 0$ we have*

$$\mathbb{P}\left(\Omega_M\right) = \mathbb{P}\left(\sup_{t\in[0,T]} \frac{1}{N} \sum_{i=1}^N \max\left\{ \left\|V_t^i\right\|_2^4, \left\|\overline{V}_t^i\right\|_2^4 \right\} \leq M \right) \geq 1 - \frac{2K}{M}, \tag{31}$$

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 [12, Lemma 3.3].

Lemma 3.10 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 propoagation of chaos through the coupling method [16, 17] using a synchronous coupling between the stochastic processes V^i and $\overline{V}^i$, see, e.g., [16, Section 4.1.2].

Proposition 3.11. *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 (3) and let $((\overline{V}_t^i)_{t\geq 0})_{i=1,\dots,N}$ be N independent copies of the strong solution to the mean-field dynamics (7). Further consider valid the assumptions of Theorem 3.2. If $(V_t^i)_{t\geq 0}$ and $(\overline{V}_t^i)_{t\geq 0}$ 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*

$$\max_{i=1,\dots,N} \sup_{t\in[0,T]} \mathbb{E}\left[\left\|V_t^i - \overline{V}_t^i\right\|_2^2 \,\middle|\, \Omega_M \right] \leq C_{\mathrm{MFA}} N^{-1} \tag{32}$$

with $C_{\mathrm{MFA}} = C_{\mathrm{MFA}}(\alpha, \lambda, \sigma, T, C_1, C_2, M, K, \mathcal{M}_2, b_1, b_2)$, where K is as in Lemma 3.10 and $\mathcal{M}_2$ denotes a second-order moment bound of ρ.

A quantitative mean-field approximation was left as an open problem in [12, Remark 3.2] due to a lack of global Lipschitz continuity of the SDE coefficients and approached since then in several steps, see Remark 1.2. While the restriction to bounded processes, which reflects the typical behavior in view of Lemma 3.10, already allows to obtain an estimate of the type (32), which is sufficient to prove convergence in probability in what follows, the recent work [31] improves (32) by firstly showing a non-probabilistic mean-field approximation, i.e., removing the necessity of conditioning on the set Ω_M as done in (32), and secondly by obtaining a pathwise estimate, see [31, Theorem 2.6]. Hence, in the light of [31], the role of the constant M can be regarded as merely an auxiliary technical tool.

Equipped with Lemma 3.10 and Proposition 3.11, we are now able to prove Theorem 3.8.

Proof of Theorem 3.8. We have the error decomposition

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N V_{K\Delta t}^i - v^*\right\|_2^2 \,\middle|\, \Omega_M\right] &\leq 3\ \underbrace{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(V_{K\Delta t}^i - V_T^i\right)\right\|_2^2 \,\middle|\, \Omega_M\right]}_{\leq C_{\mathrm{NA}}(\Delta t)^{2m} \text{ by applying classical convergence results for numerical schemes for SDEs [55]}} \\ &\quad + 3\ \underbrace{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \left(V_T^i - \overline{V}_T^i\right)\right\|_2^2 \,\middle|\, \Omega_M\right]}_{\leq C_{\mathrm{MFA}}N^{-1} \text{ using the quantitative mean-field approximation in form of Proposition 3.11}} + \frac{3}{1-\delta}\ \underbrace{\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N \overline{V}_T^i - v^*\right\|_2^2}_{\leq \mathbb{E}\left\|\overline{V}_T^1 - v^*\right\|_2^2 \leq 2\mathcal{V}(\rho_T) \leq 2\varepsilon \text{ by means of Theorem 3.7}} \\ &\leq 6C_{\mathrm{NA}}(\Delta t)^{2m} + 3C_{\mathrm{MFA}}N^{-1} + 12\varepsilon\end{aligned} \tag{33}$$

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 $K^N_{\varepsilon_{\mathrm{total}}} \subset \Omega$ the set, where (30) does not hold, we can estimate

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

where in the last step we employ Markov's inequality together with (33) to bound the first term. For the second it suffices to choose the M from (31) large enough. □

As a consequence of Theorem 3.7, the hardness of any optimization problem is necessarily encoded in the mean-field approximation. Proposition 3.11 addresses precisely this question, ensuring that, with arbitrarily high probability, the finite particle dynamics (3) keeps close to the mean-field dynamics (7). 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 (32), 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 [13, 28–30] 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 3.11. We did not precisely specify α_0 in Theorem 3.7 since it seems challenging to provide informative bounds in all generality. In Remark 4.8, 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 (24) globally for $\nu = 1/2$. For a parameter-dependent constant $c = c(\vartheta, \lambda, \sigma)$, we obtain

$$\alpha > \alpha_0 = \frac{-8}{c^2\eta^2\varepsilon} \log\left(\frac{c}{2\sqrt{2}}\,\rho_0\left(B_{\min\{R,\, c^2\eta^2\varepsilon/(8L)\}}(v^*)\right)\right) \tag{34}$$

provided that the probability mass $t \mapsto \rho_t\left(B_{c^2\eta^2\varepsilon/(8L)}(v^*)\right)$ 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 (34) 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 $d, 1/\varepsilon \to \infty$ by (34). 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 3.7. In Section 4.1 we give a proof sketch to outline the main steps. Sections 4.2–4.4 provide auxiliary results, which might be of independent interest. In Section 4.5 we complete the proof of Theorem 3.7. 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 Proof sketch

Let us first concentrate on the case $H \equiv 1$ and recall the definition of the energy functional $\mathcal{V}(\rho_t) = 1/2 \int \|v - v^*\|_2^2 \, d\rho_t(v)$. The main idea is to show that $\mathcal{V}(\rho_t)$ satisfies the differential inequality

$$\frac{d}{dt}\mathcal{V}(\rho_t) \leq -(1-\vartheta)\left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) \tag{35}$$

until the time $T = T^*$ or until $\mathcal{V}(\rho_T) \leq \varepsilon$. In the case $T = T^*$, it is easy to check that the definition of T^* in Theorem 3.7 implies $\mathcal{V}(\rho_{T^*}) \leq \varepsilon$, thus giving $\mathcal{V}(\rho_T) \leq \varepsilon$ in either case.

The first step towards (35) is to derive a differential inequality for the evolution of $\mathcal{V}(\rho_t)$ by using the dynamics of ρ. Namely, by using the (18) with test function $\phi(v) = 1/2\,\|v - v^*\|_2^2$, we derive in Lemma 4.1 the inequality

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t) \leq &-\left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) + \sqrt{2}\left(\lambda + d\sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\,\|v_\alpha(\rho_t) - v^*\|_2 \\ &+ \frac{d\sigma^2}{2}\,\|v_\alpha(\rho_t) - v^*\|_2^2 .\end{aligned} \tag{36}$$

To find suitable bounds on the second and third term in (36), we need to control the quantity $\|v_\alpha(\rho_t) - v^*\|_2$. This is achieved by the quantitative Laplace principle. Namely, under the inverse continuity property A2, Proposition 4.5 in Section 4.3 shows

$$\|v_\alpha(\rho_t) - v^*\|_2 \lesssim \ell(r) + \frac{\exp(-\alpha r)}{\rho_t(B_r(v^*))}, \quad \text{for sufficiently small } r > 0,$$

where ℓ is a strictly positive but monotonically decreasing function with $\ell(r) \to 0$ as $r \to 0$. As long as we can guarantee $\rho_t(B_r(v^*)) > 0$, the choice

$$\alpha > \frac{-\log(\rho_t(B_r(v^*))) - \log(\ell(r))}{r}$$

implies $\|v_\alpha(\rho_t) - v^*\|_2 \lesssim \ell(r)$, which can be made arbitrarily small by suitable choices of $r \ll 1$ and $\alpha \gg 1$.

Consequently, the last ingredient to the proof is to show $\rho_t(B_r(v^*)) > 0$ for all $r > 0$. To this end, we prove in Proposition 4.6 that the initial mass $\rho_0(B_r(v^*)) > 0$ can decay at most at an exponential rate for any $r > 0$, but remains strictly positive in any finite time window $[0, T]$. We emphasize that a key requirement to this result is an active Brownian motion term, i.e., $\sigma > 0$, that counteracts the deterministic movement of the drift term by inducing randomness.

The proof for the case $H \not\equiv 1$ differs slightly, because the evolution inequality (36) contains an additional term involving $\|v_\alpha(\rho_t) - v^*\|_2$, see Lemma 4.3. The idea, however, stays the same and the remaining steps of the proof can be conducted in a similar fashion.

4.2 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 $H \not\equiv 1$, respectively.

Lemma 4.1. *Let $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$, $H \equiv 1$, 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 (8). Then the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t) \leq &- \left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) + \sqrt{2}\left(\lambda + d\sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\,\|v_\alpha(\rho_t) - v^*\|_2 \\ &+ \frac{d\sigma^2}{2}\,\|v_\alpha(\rho_t) - v^*\|_2^2\,.\end{aligned} \tag{37}$$

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 (18) for all test functions in $\mathcal{C}_*^2(\mathbb{R}^d)$, see Remark 3.3. By applying (18) 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

$$T_2 \leq d\sigma^2\left(\mathcal{V}(\rho_t) + \int \|v - v^*\|_2\, d\rho_t(v)\,\|v_\alpha(\rho_t) - v^*\|_2 + \frac{1}{2}\,\|v_\alpha(\rho_t) - v^*\|_2^2\right). \tag{38}$$

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. □

Lemma 4.2. *Under the assumptions of Lemma 4.1, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt}\mathcal{V}(\rho_t) \geq -\left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) - \sqrt{2}\left(\lambda + d\sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\,\|v_\alpha(\rho_t) - v^*\|_2\,. \tag{39}$$

Proof. By following the lines of the proof of Lemma 4.1 and noticing that by Cauchy-Schwarz inequality it holds $\left\langle \int (v - v^*)\, d\rho_t(v), v_\alpha(\rho_t) - v^*\right\rangle \geq -\|\mathbb{E}(\rho_t) - v^*\|_2\,\|v_\alpha(\rho_t) - v^*\|_2$ and since moreover $\|v_\alpha(\rho_t) - v^*\|_2^2 \geq 0$, the lower bound is immediate. □

Lemma 4.3. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3 and w.l.o.g. assume $\underline{\mathcal{E}} = 0$. Let $H : \mathbb{R}^d \to [0,1]$ 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 (8). Then, provided $\max_{t\in[0,T]} \mathcal{E}(v_\alpha(\rho_t)) \leq \mathcal{E}_\infty$, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\rho_t) \leq &- \left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) + \sqrt{2}\left(\lambda + d\sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\,\|v_\alpha(\rho_t) - v^*\|_2 \\ &+ \frac{\lambda}{\eta^2}\left(L_{\mathcal{E}}\left(1 + \|v_\alpha(\rho_t) - v^*\|_2^\gamma\right)\|v_\alpha(\rho_t) - v^*\|_2\right)^{2\nu} + \frac{d\sigma^2}{2}\|v_\alpha(\rho_t) - v^*\|_2^2\,.\end{aligned}$$

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 (18) as in the proof of Lemma 4.1 yields for the evolution of $\mathcal{V}(\rho_t)$

$$\frac{d}{dt}\mathcal{V}(\rho_t) = \underbrace{-\lambda\int H^*(v)\langle v - v^*, v - v_\alpha(\rho_t)\rangle\, d\rho_t(v)}_{=:\widetilde{T}_1} + \frac{d\sigma^2}{2}\int \|v - v_\alpha(\rho_t)\|_2^2\, d\rho_t(v). \tag{40}$$

For the second term on the right-hand side, we proceed as in Equation (38). The term $\widetilde{T}_1$ on the other hand can be rewritten as

$$\begin{aligned}\widetilde{T}_1 = &-2\lambda\mathcal{V}(\rho_t) - \lambda\int H^*(v)\langle v - v^*, v^* - v_\alpha(\rho_t)\rangle\, d\rho_t(v) \\ &+ \lambda\int (1 - H^*(v))\,\|v - v^*\|_2^2\, d\rho_t(v).\end{aligned} \tag{41}$$

Let us now bound the latter two terms individually. For the second term in (41), 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 (41), 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

$$\begin{aligned}\lambda\int (1 - H^*(v))\,\|v - v^*\|_2^2\, d\rho_t(v) &\leq \lambda\int \frac{(1 - H^*(v))}{\eta^2}\mathcal{E}(v)^{2\nu}\, d\rho_t(v) \leq \frac{\lambda}{\eta^2}\mathcal{E}(v_\alpha(\rho_t))^{2\nu} \\ &\leq \frac{\lambda}{\eta^2}\left(L_{\mathcal{E}}\left(1 + \|v_\alpha(\rho_t) - v^*\|_2^\gamma\right)\|v_\alpha(\rho_t) - v^*\|_2\right)^{2\nu},\end{aligned}$$

where the last step used A3. Employing the last two inequalities in (41) and inserting the result together with (38) into (40), gives the result. □

Lemma 4.4. *Under the assumptions of Lemma 4.3, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt}\mathcal{V}(\rho_t) \geq -\left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) - \sqrt{2}\left(\lambda + d\sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\,\|v_\alpha(\rho_t) - v^*\|_2\,. \tag{42}$$

Proof. Analogously to the proof of Lemma 4.2, by following the lines of the proof of Lemma 4.3 and noticing that for $\widetilde{T}_1$ 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. □

4.3 Quantitative Laplace principle

The Laplace principle (6) asserts that $-\log(\|\omega_\alpha\|_{L_1(\varrho)})/\alpha \to \underline{\mathcal{E}}$ as $\alpha \to \infty$ 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 $\underline{\mathcal{E}}$, 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 4.5. *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 $\widetilde{r} \geq r > 0$. 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

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \int_{B_{\widetilde{r}}(v^*)} \|v - v^*\|_2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\varrho)}} \, d\varrho(v) + \int_{(B_{\widetilde{r}}(v^*))^c} \|v - v^*\|_2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\varrho)}} \, d\varrho(v).$$

The first term is bounded by $\widetilde{r}$ since $\|v - v^*\|_2 \leq \widetilde{r}$ for all $v \in B_{\widetilde{r}}(v^*)$. 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

$$\begin{aligned}\int_{(B_{\widetilde{r}}(v^*))^c} \|v - v^*\|_2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\varrho)}} \, d\varrho(v) &\leq \frac{1}{\exp(-\alpha\mathcal{E}_r)\varrho(B_r(v^*))} \int_{(B_{\widetilde{r}}(v^*))^c} \|v - v^*\|_2 \, \omega_\alpha(v) \, d\varrho(v) \\ &\leq \frac{\exp\left(-\alpha\left(\inf_{v \in (B_{\widetilde{r}}(v^*))^c} \mathcal{E}(v) - \mathcal{E}_r\right)\right)}{\varrho(B_r(v^*))} \int \|v - v^*\|_2 \, d\varrho(v).\end{aligned}$$

Thus, for any $\widetilde{r} \geq r > 0$ we obtain

$$\|v_\alpha(\varrho) - v^*\|_2 \leq \widetilde{r} + \frac{\exp\left(-\alpha\left(\inf_{v \in (B_{\widetilde{r}}(v^*))^c} \mathcal{E}(v) - \mathcal{E}_r\right)\right)}{\varrho(B_r(v^*))} \int \|v - v^*\|_2 \, d\varrho(v). \tag{43}$$

Let us now choose $\widetilde{r} = (q + \mathcal{E}_r)^\nu/\eta$. This choice satisfies $\widetilde{r} \leq \mathcal{E}_\infty^\nu/\eta$ by the assumption $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, and furthermore $\widetilde{r} \geq r$, since A2 with $\underline{\mathcal{E}} = 0$ and $r \leq R_0$ implies

$$\widetilde{r} = \frac{(q + \mathcal{E}_r)^\nu}{\eta} \geq \frac{\mathcal{E}_r^\nu}{\eta} = \frac{\left(\sup_{v \in B_r(v^*)} \mathcal{E}(v)\right)^\nu}{\eta} \geq \sup_{v \in B_r(v^*)} \|v - v^*\|_2 = r.$$

Thus, using again A2 with $\underline{\mathcal{E}} = 0$, $\inf_{v \in (B_{\widetilde{r}}(v^*))^c} \mathcal{E}(v) - \mathcal{E}_r \geq \min\left\{\mathcal{E}_\infty, (\eta\widetilde{r})^{1/\nu}\right\} - \mathcal{E}_r = (\eta\widetilde{r})^{1/\nu} - \mathcal{E}_r = q$. Inserting this and the definition of $\widetilde{r}$ into (43), we obtain the result. □

4.4 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 $\phi_r : \mathbb{R}^d \to [0,1]$ with $\operatorname{supp}\phi_r = B_r(v^*)$ according to

$$\phi_r(v) := \begin{cases} \exp\left(1 - \frac{r^2}{r^2 - \|v-v^*\|_2^2}\right), & \text{if } \|v-v^*\|_2 < r, \\ 0, & \text{else,} \end{cases} \tag{44}$$

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 3.1, since $\phi_r \in \mathcal{C}_c^\infty(\mathbb{R}^d)$.

To do so, we compute the derivatives

$$\nabla\phi_r(v) = -2r^2 \frac{v - v^*}{\left(r^2 - \|v-v^*\|_2^2\right)^2} \phi_r(v), \tag{45}$$

$$\Delta\phi_r(v) = 2r^2 \left(\frac{2\left(2\|v-v^*\|_2^2 - r^2\right)\|v-v^*\|_2^2 - d\left(r^2 - \|v-v^*\|_2^2\right)^2}{\left(r^2 - \|v-v^*\|_2^2\right)^4} \right) \phi_r(v). \tag{46}$$

Proposition 4.6. *Let $H : \mathbb{R} \to [0,1]$ 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 (8) in the sense of Definition 3.1 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*

$$\rho_t\left(B_r(v^*)\right) \geq \left(\int \phi_r(v)\, d\rho_0(v) \right) \exp\left(-pt\right), \quad \textit{where} \tag{47}$$

$$p := \max\left\{ \frac{2\lambda(\sqrt{c}r + B)\sqrt{c}}{(1-c)^2 r} + \frac{2\sigma^2(cr^2 + B^2)(2c+d)}{(1-c)^4 r^2}, \frac{4\lambda^2}{(2c-1)\sigma^2} \right\} \tag{48}$$

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

$$(2c-1)c \geq d(1-c)^2. \tag{49}$$

Remark 4.7. Let us comment in what follows on two technical details of Proposition 4.6.

(i) Note that neither the definition of $v_\alpha(\rho_t)$ nor v^* play a significant role in the proof. The same result holds for an arbitrary $v \in \mathbb{R}^d$ so that $\sup_{t\in[0,T]} \|v_\alpha(\rho_t) - v\|_2 \leq B < \infty$ and for arbitrary continuous maps $u \in \mathcal{C}([0,T], \mathbb{R}^d)$ to replace $v_\alpha(\rho_t)$ as long as ρ weakly solves the Fokker-Planck equation (8) with u in place of $v_\alpha(\rho_t)$.

(ii) In case the reader may have wondered about the crucial role of the stochastic terms in (1) and (3), or the diffusion in the macroscopic models (7) and (8), Proposition 4.6 precisely explains where positive diffusion $\sigma > 0$ is actually used to ensure mass around the minimizer v^* (compare Proposition 4.5). We require $\sigma > 0$ in Proposition 4.6 to ensure a finite decay rate $q < \infty$, see the definition in Equation (47). Intuitively, we can understand the measure ρ_t as having a deterministic component, which evolves according to the drift term in the Fokker-Planck equation (8) and whose associated mass may leave $B_r(v^*)$ 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 in the case $\sigma = 0$.

Proof of Proposition 4.6. By definition of the mollifier ϕ_r in (44) we have $0 \leq \phi_r(v) \leq 1$ and $\mathrm{supp}(\phi_r) = B_r(v^*)$. This implies

$$\rho_t\left(B_r(v^*)\right) = \rho_t\left(\left\{v \in \mathbb{R}^d : \|v - v^*\|_2 \leq r\right\}\right) \geq \int \phi_r(v)\, d\rho_t(v). \tag{50}$$

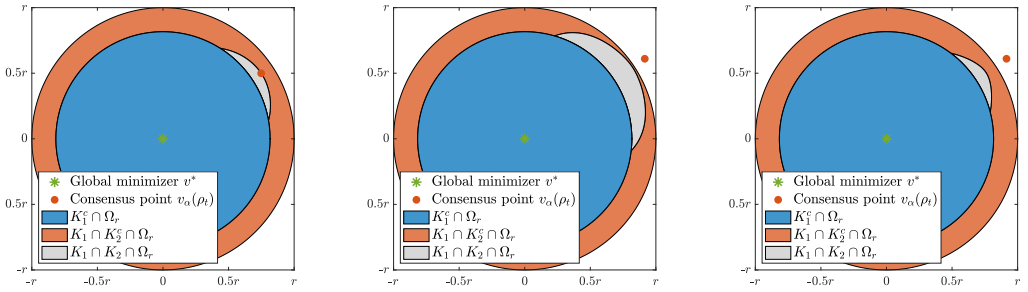
Our strategy is to derive a lower bound for the right-hand side of this inequality. Using the weak solution property of ρ and the fact that $\phi_r \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, we obtain

$$\frac{d}{dt}\int \phi_r(v)\, d\rho_t(v) = \int \left(T_1(v) + T_2(v)\right) d\rho_t(v) \tag{51}$$

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 (48) 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 $K_1 := \left\{v \in \mathbb{R}^d : \|v - v^*\|_2 > \sqrt{c}r\right\}$ and

$$K_2 := \Bigg\{v \in \mathbb{R}^d : -\lambda H^*(v)\langle v - v_\alpha(\rho_t), v - v^*\rangle \left(r^2 - \|v - v^*\|_2^2\right)^2 > \tilde{c}r^2\frac{\sigma^2}{2}\|v - v_\alpha(\rho_t)\|_2^2\,\|v - v^*\|_2^2\Bigg\},$$

where c adheres to (49), 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.



(a) Decomposition for $v_\alpha(\rho_t) \in \Omega_r$ and $\sigma = 0.2$

(b) Decomposition for $v_\alpha(\rho_t) \notin \Omega_r$ and $\sigma = 0.2$

(c) Decomposition for $v_\alpha(\rho_t) \notin \Omega_r$ and $\sigma = 1$

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 4.6 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 $-4\lambda^2/((2c-1)\sigma^2)$.

In the following we treat each of these three subsets separately.

Subset $K_1^c \cap \Omega_r$: We have $\|v - v^*\|_2 \leq \sqrt{c}r$ 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 (44), for T_1 we

get by using Cauchy-Schwarz inequality and $H^* \leq 1$

$$T_1(v) = -\lambda H^*(v)\langle v - v_\alpha(\rho_t), \nabla\phi_r(v)\rangle = -\lambda H^*(v)\left\langle v - v_\alpha(\rho_t), \frac{-2r^2(v-v^*)\phi_r(v)}{\left(r^2 - \|v-v^*\|_2^2\right)^2}\right\rangle$$
$$\geq -2r^2\lambda\frac{\|v - v_\alpha(\rho_t)\|_2\,\|v-v^*\|_2}{\left(r^2-\|v-v^*\|_2^2\right)^2}\phi_r(v) \geq -\frac{2\lambda(\sqrt{c}r+B)\sqrt{c}}{(1-c)^2 r}\phi_r(v) =: -p_1\phi_r(v),$$

where the last bound is due to $\|v - v_\alpha(\rho_t)\|_2 \leq \|v-v^*\|_2 + \|v^* - v_\alpha(\rho_t)\|_2 \leq \sqrt{c}r + B$. Similarly, by computing $\Delta\phi_r$ and inserting it, for T_2 we obtain

$$T_2(v) = \sigma^2 r^2\,\|v - v_\alpha(\rho_t)\|_2^2\,\frac{2\left(2\|v-v^*\|_2^2 - r^2\right)\|v-v^*\|_2^2 - d\left(r^2-\|v-v^*\|_2^2\right)^2}{\left(r^2-\|v-v^*\|_2^2\right)^4}\phi_r(v)$$
$$\geq -\frac{2\sigma^2(cr^2+B^2)(2c+d)}{(1-c)^4 r^2}\phi_r(v) =: -p_2\phi_r(v),$$

where we used $\|v - v_\alpha(\rho_t)\|_2^2 \leq 2\big(\|v-v^*\|_2^2 + \|v^* - v_\alpha(\rho_t)\|_2^2\big) \leq 2(cr^2 + B^2)$.
Subset $K_1 \cap K_2^c \cap \Omega_r$: By the definition of K_1 and K_2 we have $\|v - v^*\|_2 > \sqrt{c}r$ and

$$-\lambda H^*(v)\langle v - v_\alpha(\rho_t), v - v^*\rangle\left(r^2 - \|v-v^*\|_2^2\right)^2 \leq \tilde{c}r^2\frac{\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\,\|v-v^*\|_2^2\,. \tag{52}$$

Our goal now is to show $T_1(v) + T_2(v) \geq 0$ for all v in this subset. We first compute

$$\frac{T_1(v)+T_2(v)}{2r^2\phi_r(v)} = \lambda H^*(v)\frac{\langle v - v_\alpha(\rho_t), v-v^*\rangle\left(r^2-\|v-v^*\|_2^2\right)^2}{\left(r^2-\|v-v^*\|_2^2\right)^4}$$
$$+\frac{\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\,\frac{2\left(2\|v-v^*\|_2^2 - r^2\right)\|v-v^*\|_2^2 - d\left(r^2-\|v-v^*\|_2^2\right)^2}{\left(r^2-\|v-v^*\|_2^2\right)^4}.$$

Therefore we have $T_1(v) + T_2(v) \geq 0$ whenever we can show

$$\left(-\lambda H^*(v)\langle v - v_\alpha(\rho_t), v-v^*\rangle + \frac{d\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\right)\left(r^2-\|v-v^*\|_2^2\right)^2$$
$$\leq \sigma^2\,\|v-v_\alpha(\rho_t)\|_2^2\left(2\|v-v^*\|_2^2 - r^2\right)\|v-v^*\|_2^2\,. \tag{53}$$

Now note that the first summand on the left-hand side in (53) can be upper bounded by means of Condition (52) and by using the relation $\tilde{c} = 2c - 1$. More precisely,

$$-\lambda H^*(v)\langle v-v_\alpha(\rho_t), v-v^*\rangle\left(r^2-\|v-v^*\|_2^2\right)^2 \leq \tilde{c}r^2\frac{\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\,\|v-v^*\|_2^2$$
$$= (2c-1)r^2\frac{\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\,\|v-v^*\|_2^2 \leq \left(2\|v-v^*\|_2^2 - r^2\right)\frac{\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\,\|v-v^*\|_2^2\,,$$

where the last inequality follows since $v \in K_1$. For the second term on the left-hand side in (53) we can use $d(1-c)^2 \leq (2c-1)c$ as per (49), to get

$$\frac{d\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\left(r^2-\|v-v^*\|_2^2\right)^2 \leq \frac{d\sigma^2}{2}\,\|v-v_\alpha(\rho_t)\|_2^2\,(1-c)^2 r^4$$
$$\leq \frac{\sigma^2}{2}\|v-v_\alpha(\rho_t)\|_2^2\,(2c-1)r^2cr^2 \leq \frac{\sigma^2}{2}\|v-v_\alpha(\rho_t)\|_2^2\left(2\|v-v^*\|_2^2 - r^2\right)\|v-v^*\|_2^2\,.$$

Hence, (53) 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 $\|v - v^*\|_2 > \sqrt{c}r$ and

$$-\lambda H^*(v)\langle v - v_\alpha(\rho_t), v - v^*\rangle \left(r^2 - \|v - v^*\|_2^2\right)^2 > \tilde{c}r^2 \frac{\sigma^2}{2} \|v - v_\alpha(\rho_t)\|_2^2 \|v - v^*\|_2^2 . \tag{54}$$

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 (54) 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 $\|v - v^*\|_2 \geq \sqrt{c}r$, provided c satisfies $2(2c - 1)c \geq (1 - c)^2 d$ as implied by (49).
Concluding the proof: Using the evolution of ϕ_r as in (51), 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 (50). □

4.5 Proof of Theorem 3.7

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 3.7 when $H \equiv 1$. W.l.o.g. we may assume $\underline{\mathcal{E}} = 0$. Let us first choose the parameter α such that

$$\begin{aligned}\alpha > \alpha_0 := \frac{1}{q_\varepsilon}\Bigg(\log\left(\frac{4\sqrt{2\mathcal{V}(\rho_0)}}{c\,(\vartheta, \lambda, \sigma)\,\sqrt{\varepsilon}}\right) + \frac{p}{(1 - \vartheta)\,(2\lambda - d\sigma^2)} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) \\ - \log \rho_0\big(B_{r_\varepsilon/2}(v^*)\big)\Bigg),\end{aligned} \tag{55}$$

where we introduce the definitions

$$c(\vartheta,\lambda,\sigma) := \min\left\{\frac{\vartheta}{2}\frac{\left(2\lambda - d\sigma^2\right)}{\sqrt{2}\left(\lambda + d\sigma^2\right)}, \sqrt{\vartheta\frac{(2\lambda - d\sigma^2)}{d\sigma^2}}\right\} \tag{56}$$

as well as

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

Moreover, p is as defined in (48) in Proposition 4.6 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 $r_\varepsilon \le R_0$. Furthermore, recalling the notation $\mathcal{E}_r = \sup_{v\in B_r(v^*)}\mathcal{E}(v)$ from Proposition 4.5, we have $q_\varepsilon + \mathcal{E}_{r_\varepsilon} \le 2q_\varepsilon \le \mathcal{E}_\infty$ 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 $\mathcal{E}(v) \le q_\varepsilon$ for all $v \in B_{s_{q_\varepsilon}}(v^*)$, thus yielding also $r_\varepsilon > 0$.

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

$$T_\alpha := \sup\left\{t \ge 0 : \mathcal{V}(\rho_{t'}) > \varepsilon \text{ and } \|v_\alpha(\rho_{t'}) - v^*\|_2 < C(t') \text{ for all } t' \in [0,t]\right\} \tag{57}$$

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 $T_\alpha \in \left[\frac{1-\vartheta}{(1+\vartheta/2)}T^*, T^*\right]$ 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 3.2 and $t \mapsto \|v_\alpha(\rho_t) - v^*\|_2$ being continuous due to [12, 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 4.5 with q_ε and r_ε gives the latter since

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

where the first inequality in the last line holds by the choice of α in (55).

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 4.1 provides an upper bound on $\frac{d}{dt}\mathcal{V}(\rho_t)$ given by

$$\begin{aligned}
\frac{d}{dt}\mathcal{V}(\rho_t) \le &-\left(2\lambda - d\sigma^2\right)\mathcal{V}(\rho_t) + \sqrt{2}\left(\lambda + d\sigma^2\right)\sqrt{\mathcal{V}(\rho_t)}\,\|v_\alpha(\rho_t) - v^*\|_2 \\
&+ \frac{d\sigma^2}{2}\|v_\alpha(\rho_t) - v^*\|_2^2.
\end{aligned} \tag{58}$$

Combining this with the definition of T_α in (57) we have by construction

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

Analogously, for (ii), by Lemma 4.2, we obtain a lower bound on $\frac{d}{dt}\mathcal{V}(\rho_t)$ of the form

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

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

$$\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0)\exp\left(-(1-\vartheta)\left(2\lambda - d\sigma^2\right)t\right), \tag{59}$$

$$\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0)\exp\left(-(1+\vartheta/2)\left(2\lambda - d\sigma^2\right)t\right), \tag{60}$$

thereby proving (i) and (ii). We further note that the definition of T_α in (57) together with the definition of $C(t)$ and (59) permits to control

$$\max_{t\in[0,T_\alpha]} \|v_\alpha(\rho_t) - v^*\|_2 \leq \max_{t\in[0,T_\alpha]} C(t) \leq C(0). \tag{61}$$

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

Case $T_\alpha \geq T^*$: We can use the definition of T^* in (27) and the time-evolution bound of $\mathcal{V}(\rho_t)$ in (59) to conclude that $\mathcal{V}(\rho_{T^*}) \leq \varepsilon$. Hence, by definition of T_α in (57) 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 (57), $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$. Thus, $\varepsilon = \mathcal{V}(\rho_{T_\alpha}) \geq \mathcal{V}(\rho_0)\exp\left(-(1+\vartheta/2)\left(2\lambda - d\sigma^2\right)T_\alpha\right)$ by (60), which can be reordered as

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

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 (55). 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 (57) itself. To this end, by applying again Proposition 4.5 with q_ε and r_ε, and recalling that $\varepsilon < \mathcal{V}(\rho_{T_\alpha})$, we get

$$\begin{aligned}\|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 &\leq \frac{(q_\varepsilon + \mathcal{E}_{r_\varepsilon})^\nu}{\eta} + \frac{\exp\left(-\alpha q_\varepsilon\right)}{\rho_{T_\alpha}\left(B_{r_\varepsilon}(v^*)\right)}\int \|v - v^*\|_2\, d\rho_{T_\alpha}(v) \\ &< \frac{c\,(\vartheta,\lambda,\sigma)\sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{\exp\left(-\alpha q_\varepsilon\right)}{\rho_{T_\alpha}\left(B_{r_\varepsilon}(v^*)\right)}\sqrt{2\mathcal{V}(\rho_{T_\alpha})}.\end{aligned} \tag{62}$$

Since, thanks to (61), 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 4.6 guarantees that there exists a $p > 0$ not depending on α (but depending on B and r_ε) with

$$\rho_{T_\alpha}(B_{r_\varepsilon}(v^*)) \geq \left(\int \phi_{r_\varepsilon}(v)\, d\rho_0(v)\right)\exp(-pT_\alpha) \geq \frac{1}{2}\,\rho_0\left(B_{r_\varepsilon/2}(v^*)\right)\exp(-pT^*) > 0,$$

where we used $v^* \in \operatorname{supp}(\rho_0)$ for bounding the initial mass ρ_0 and the fact that ϕ_r (as defined in Equation (44)) is bounded from below on $B_{r/2}(v^*)$ by $1/2$. With this we can continue the chain of inequalities in (62) to obtain

$$\begin{aligned}\|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 &< \frac{c\,(\vartheta,\lambda,\sigma)\sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2\exp\left(-\alpha q_\varepsilon\right)}{\rho_0\left(B_{r_\varepsilon/2}(v^*)\right)\exp(-pT^*)}\sqrt{2\mathcal{V}(\rho_{T_\alpha})} \\ &\leq c\,(\vartheta,\lambda,\sigma)\sqrt{\mathcal{V}(\rho_{T_\alpha})} = C(T_\alpha),\end{aligned} \tag{63}$$

where the first inequality in the last line holds by the choice of α in (55). This establishes the desired contradiction, again as consequence of the continuity of the mappings $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \|v_\alpha(\rho_t) - v^*\|_2$. □

We now consider the case of an active cutoff function H with $H(x)=1$ whenever $x\geq 0$.

Proof of Theorem 3.7 when $H \not\equiv 1$. The proof follows the lines of the one for the inactive cutoff $H \equiv 1$, but requires some modifications since Lemmas 4.1 and 4.2 need to be replaced by Lemmas 4.3 and 4.4, 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

$$\alpha > \alpha_0 := \frac{1}{q_\varepsilon}\left(\log\left(\frac{4\sqrt{2\mathcal{V}(\rho_0)}}{C_\varepsilon}\right) + \frac{p}{(1-\vartheta)\,(2\lambda - d\sigma^2)} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) - \log \rho_0\big(B_{r_\varepsilon/2}(v^*)\big)\right), \tag{64}$$

where C_ε is obtained when replacing with ε each $\mathcal{V}(\rho_t)$ in $C(t)$ defined as

$$C(t) := \min\left\{ \frac{\mathcal{E}_\infty}{2L_\mathcal{E}}, \left(\frac{\mathcal{E}_\infty}{2L_\mathcal{E}}\right)^{1/(1+\gamma)}, \frac{\vartheta}{4}\frac{(2\lambda - d\sigma^2)}{\sqrt{2}\,(\lambda + d\sigma^2)}\sqrt{\mathcal{V}(\rho_t)}, \sqrt{\frac{\vartheta}{2}\frac{(2\lambda - d\sigma^2)}{d\sigma^2}\mathcal{V}(\rho_t)}, \left(\frac{\vartheta}{4}\frac{\eta^2}{L_\mathcal{E}^{2\nu}}\frac{(2\lambda - d\sigma^2)}{\lambda}\mathcal{V}(\rho_t)\right)^{1/(2\nu)}, \left(\frac{\vartheta}{4}\frac{\eta^2}{L_\mathcal{E}^{2\nu}}\frac{(2\lambda - d\sigma^2)}{\lambda}\mathcal{V}(\rho_t)\right)^{1/(2\nu(1+\gamma))} \right\}. \tag{65}$$

Moreover, r_ε is as defined before, p as in (48) 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 (57), however with the modified definition of $C(t)$ from (65). It is straightforward to check that $T_\alpha > 0$ by choice of α in (64). 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 (65) it holds

$$\max_{t\in[0,T_\alpha]} \mathcal{E}(v_\alpha(\rho_t)) \leq \max_{t\in[0,T_\alpha]} L_\mathcal{E}(1 + \|v_\alpha(\rho_t) - v^*\|_2^\gamma)\,\|v_\alpha(\rho_t) - v^*\|_2 \leq \mathcal{E}_\infty, \tag{66}$$

Lemmas 4.3 and 4.4 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 (65), 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 (64) allows to derive the contradiction $\|v_\alpha(\rho_{T_\alpha}) - v^*\|_2 < C(T_\alpha)$ by employing Propositions 4.5 and 4.6. □

Remark 4.8 (Informal lower bound for α_0)**.** As mentioned in Section 3.3, insightful lower bounds on the required α_0 in Theorem 3.7 may be interesting in view of better understanding the convergence of the microscopic system (3) to the mean-field limit (7). 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 (58) in the proof of Theorem 3.7, α should be large enough to ensure

$$\|v_\alpha(\rho_t) - v^*\|_2 \leq c\,(\vartheta, \lambda, \sigma)\,\sqrt{\mathcal{V}(\rho_t)} \quad \text{for all } t \in [0, T], \tag{67}$$

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 4.5 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

$$\alpha_0 := \sup_{t \in [0,T]} \frac{-8}{c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon} \log\left(\frac{c(\vartheta, \lambda, \sigma)}{2\sqrt{2}} \rho_t\Big(B_{\min\{R,\, c(\vartheta,\lambda,\sigma)^2 \eta^2 \varepsilon / (8L)\}}(v^*) \Big) \right), \tag{68}$$

suggesting that α_0 is strongly related to the time-evolution of the probability mass of ρ_t around v^*. Recalling Proposition 4.6, 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 (68) with the corresponding p as in (48) leads to overly stringent requirements on α_0, which are reflected by the respective second summands in (55) and (64). 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 (68) becomes

$$\alpha_0 = \frac{-8}{c(\vartheta, \lambda, \sigma)^2 \eta^2 \varepsilon} \log\left(\frac{c(\vartheta, \lambda, \sigma)}{2\sqrt{2}} \rho_0\Big(B_{\min\{R,\, c(\vartheta,\lambda,\sigma)^2 \eta^2 \varepsilon / (8L)\}}(v^*) \Big) \right). \tag{69}$$

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 3.11. After giving the proof of the auxiliary Lemma 3.10, which ensures that the dynamics is to some extent bounded, we prove Proposition 3.11.

Proof of Lemma 3.10. By combining the ideas of [12, Lemma 3.4] with a Doob-like inequality, we derive a bound for $\mathbb{E} \sup_{t \in [0,T]} \frac{1}{N} \sum_{i=1}^N \max\left\{ \|V_t^i\|_2^4 + \|\overline{V}_t^i\|_2^4 \right\}$, which ensures that $\widehat{\rho}_t^N, \bar{\rho}_t^N \in \mathcal{P}_4(\mathbb{R}^d)$ with high probability. Here, $\bar{\rho}^N$ denotes the empirical measure associated with the processes $(\overline{V}^i)_{i=1,\dots,N}$.
Employing standard inequalities shows

$$\begin{aligned} \mathbb{E} \sup_{t \in [0,T]} \left\| V_t^i \right\|_2^4 \lesssim \mathbb{E} \left\| V_0^i \right\|_2^4 &+ \lambda^4 \, \mathbb{E} \sup_{t \in [0,T]} \left\| \int_0^t \left(V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N) \right) d\tau \right\|_2^4 \\ &+ \sigma^4 \, \mathbb{E} \sup_{t \in [0,T]} \left\| \int_0^t \left\| V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N) \right\|_2 dB_\tau^i \right\|_2^4, \end{aligned} \tag{70}$$

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 [53, Corollary 3.2.6] combined with the regularity established in [12, Lemma 3.4]. This allows to apply the Burkholder-Davis-Gundy inequality [58, Chapter IV, Theorem 4.1], which yields

$$\mathbb{E} \sup_{t \in [0,T]} \left\| \int_0^t \left\| V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N) \right\|_2 dB_\tau^i \right\|_2^4 \lesssim \mathbb{E} \left(\int_0^T \left\| V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N) \right\|_2^2 d\tau \right)^2.$$

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 (70) by means of Jensen's inequality and utilizing [12, Lemma 3.3] yields

$$\mathbb{E} \sup_{t\in[0,T]} \left\|V_t^i\right\|_2^4 \leq C \left(1 + \mathbb{E} \left\|V_0^i\right\|_2^4 + \mathbb{E} \int_0^T \left\|V_\tau^i\right\|_2^4 + \int \|v\|_2^4 \, d\widehat{\rho}_\tau^N(v) \, d\tau\right) \tag{71}$$

with a constant $C = C(\lambda, \sigma, d, T, b_1, b_2)$. Averaging (71) over i allows to bound

$$\mathbb{E} \sup_{t\in[0,T]} \int \|v\|_2^4 \, d\widehat{\rho}_t^N(v) \leq C \left(1 + \mathbb{E} \int \|v\|_2^4 \, d\widehat{\rho}_0^N(v) + 2 \int_0^T \mathbb{E} \sup_{\widehat{\tau}\in[0,\tau]} \int \|v\|_2^4 \, d\widehat{\rho}_{\widehat{\tau}}^N(v) \, d\tau\right),$$

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 we can show $\mathbb{E} \sup_{t\in[0,T]} \int \|v\|_2^4 \, d\bar{\rho}_t^N(v) \leq K$. Equation (31) follows now from Markov's inequality. □

We now present the proof of Proposition 3.11.

Proof of Proposition 3.11. By exploiting the boundedness thanks to Lemma 3.10 through a cutoff technique, we can follow the steps taken in [27, Theorem 3.1].
Let us define the cutoff function

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

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

$$\mathbb{E} \left\|V_t^i - \overline{V}_t^i\right\|_2^2 I_M(t) \lesssim c \int_0^t \mathbb{E} \left(\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 + \left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2\right) I_M(\tau) \, d\tau \tag{72}$$

for $c = \left(\lambda^2 T + \sigma^2\right)$. Here we directly used that the processes V_t^i and $\overline{V}_t^i$ 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 $\overline{V}_\tau^i$. Then, by using the same arguments as in the proofs of [12, Lemma 3.2] and [27, Lemma 3.1] with the care of taking into consideration the multiplication with the random variable $I_M(\tau)$, we obtain

$$\begin{aligned} \mathbb{E}\left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2 I_M(\tau) &\lesssim \mathbb{E}\left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\right\|_2^2 I_M(\tau) + \mathbb{E}\left\|v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2 I_M(\tau) \\ &\leq C \left(\max_{i=1,\dots,N} \mathbb{E} \left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 I_M(\tau) + N^{-1}\right) \end{aligned}$$

for a constant $C = C(\alpha, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. After plugging the latter into (72) and taking the maximum over i, the quantitative mean-field approximation result (32) 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 $t \in [0,T]$. □

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 [12, 28, 33, 34], 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 (1), 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 [27, 28], polarized CBO adjusted to identify multiple minimizers simultaneously [11], as well as related metaheuristics including, for instance, Particle Swarm Optimization [32, 40, 44], which can be regarded as a second-order variant of CBO with inertia [22, 32]. For CBO with anisotropic Brownian motions, which are especially relevant in high-dimensional optimization problems [13], for CBO with memory effects and gradient information, which can be beneficial in signal processing and machine learning applications [15, 59], for CBO reconfigured for multi-objective optimization, as well as for constrained CBO, this has already been done in [30], [59], [9], and [10], 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] L. Arnold. *Stochastic differential equations: Theory and applications.* Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974. Translated from the German.

[5] T. Bäck, D. B. Fogel, and Z. Michalewicz, editors. *Handbook of evolutionary computation.* Institute of Physics Publishing, Bristol; Oxford University Press, New York, 1997.

[6] 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. *J. Open Source Softw.*, 9(98):6611, 2024.

[7] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*, 35(3):268–308, 2003.

[8] 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.

[9] 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.

[10] G. Borghi, M. Herty, and L. Pareschi. Constrained consensus-based optimization. *SIAM J. Optim.*, 33(1):211–236, 2023.

[11] L. Bungert, T. Roith, and P. Wacker. Polarized consensus-based dynamics for optimization and sampling. *Math. Program.*, pages 1–31, 2024.

[12] 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.

[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. *J. Mach. Learn. Res.*, 25:Paper No. 214, 1–51, 2024.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] L. Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *Trans. Mach. Learn. Res.*, 2022, 2022.

[21] 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.

[22] 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.

[23] 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.

[24] J. Dong and X. T. Tong. Replica exchange for non-convex optimization. *J. Mach. Learn. Res.*, 22:Paper No. 173, 1–59, 2021.

[25] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

[26] D. B. Fogel. *Evolutionary computation. Toward a new philosophy of machine intelligence*. IEEE Press, Piscataway, NJ, second edition, 2000.

[27] 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.

[28] 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, 1–55, 2021.

[29] 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.

[30] 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.

[31] N. J. Gerber, F. Hoffmann, and U. Vaes. Mean-field limits for consensus-based optimization and sampling. *arXiv preprint arXiv:2312.07373*, 2023.

[32] 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.

[33] 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.

[34] 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.

[35] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[36] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001.

[37] 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.

[38] 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.

[39] H. Huang and J. Qiu. On the mean-field limit for the consensus-based optimization. *Math. Methods Appl. Sci.*, 45(12):7814–7831, 2022.

[40] 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.

[41] 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.

[42] 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.

[43] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz 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.

[44] 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.

[45] 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.

[46] Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database, 2010.

[47] 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.

[48] J. Lu, E. Tadmor, and A. Zenginoglu. Swarm-based gradient descent method for non-convex optimization. *arXiv preprint arXiv:2211.17157*, 2022.

[49] 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.

[50] 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.

[51] P. D. Miller. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.

[52] 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.

[53] B. Øksendal. *Stochastic differential equations: An introduction with applications*. Springer-Verlag, Berlin, sixth edition, 2003.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] K. Riedl. Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law. *Eur. J. Appl. Math.*, pages 1–32, 2023.

[60] 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.

[61] K. Riedl, T. Klock, C. Geldhauser, and M. Fornasier. How Consensus-Based Optimization can be Interpreted as a Stochastic Relaxation of Gradient Descent. *ICML Workshop Differentiable Almost Everything: Differentiable Relaxations, Algorithms, Operators, and Simulators*, 2024.

[62] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[63] 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.

[64] 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.

[65] F. J. Solis and R. J.-B. Wets. Minimization by random search techniques. *Math. Oper. Res.*, 6(1):19–30, 1981.

[66] 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.

[67] C. Totzeck and M.-T. Wolfram. Consensus-based global optimization with personal best. *Math. Biosci. Eng.*, 17(5):6026–6044, 2020.

[68] F. van den Bergh. *An analysis of particle swarm optimizers.* PhD thesis, University of Pretoria, 2007.

[69] 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 3.10. By combining the ideas of [12, Lemma 3.4] with a Doob-like inequality, we derive a bound for $\mathbb{E}\sup_{t\in[0,T]} \frac{1}{N}\sum_{i=1}^N \max\left\{\|V_t^i\|_2^4, \|\overline{V}_t^i\|_2^4\right\}$, which ensures that $\widehat{\rho}_t^N, \bar{\rho}_t^N \in \mathcal{P}_4(\mathbb{R}^d)$ with high probability. Here, $\bar{\rho}^N$ denotes the empirical measure associated with the processes $(\overline{V}^i)_{i=1,\dots,N}$. For notational simplicity, but without loss of generality, we restrict ourselves to the case $H \equiv 1$ in what follows.
By employing the inequality $(x+y)^q \le 2^{q-1}(x^q+y^q)$, $q \ge 1$ we note that

$$\begin{aligned}\left\|V_t^i\right\|_2^{2p} \le 2^{2p-1}\left\|V_0^i\right\|_2^{2p} &+ 2^{2(2p-1)}\lambda^{2p}\left\|\int_0^t \left(V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right)d\tau\right\|_2^{2p} \\ &+ 2^{2(2p-1)}\sigma^{2p}\left\|\int_0^t \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2 dB_\tau^i\right\|_2^{2p}\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,T]}\left\|V_t^i\right\|_2^{2p} \le 2^{2p-1}\mathbb{E}\left\|V_0^i\right\|_2^{2p} &+ 2^{2(2p-1)}\lambda^{2p}\mathbb{E}\sup_{t\in[0,T]}\left\|\int_0^t \left(V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right)d\tau\right\|_2^{2p} \\ &+ 2^{2(2p-1)}\sigma^{2p}\mathbb{E}\sup_{t\in[0,T]}\left\|\int_0^t \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2 dB_\tau^i\right\|_2^{2p}.\end{aligned} \tag{73}$$

The second term on the right-hand side of (73) can be further bounded by

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,T]}\left\|\int_0^t \left(V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right)d\tau\right\|_2^{2p} &\le \max\{1,T^{2p-1}\}\mathbb{E}\sup_{t\in[0,T]}\int_0^t \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^{2p} d\tau \\ &\le \max\{1,T^{2p-1}\}\mathbb{E}\int_0^T \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^{2p} d\tau\end{aligned} \tag{74}$$

as a consequence of Jensen's inequality. For the third term on the right-hand side of (73) we first note that the expression $\int_0^t \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2 dB_\tau^i$ is a martingale. This is due to [53, Corollary 3.2.6] since its expected quadratic variation is finite as required by [53, Definition 3.1.4]. The latter immediately follows from the regularity established in [12, Lemma 3.4]. Therefore we can apply the Burkholder-Davis-Gundy inequality [58, Chapter IV, Theorem 4.1], which gives for a generic constant C_{2p} depending only on the dimension d the bound

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,T]}\left\|\int_0^t \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2 dB_\tau^i\right\|_2^{2p} &\le C_{2p}\sup_{t\in[0,T]}\mathbb{E}\left(\int_0^t \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^2 d\tau\right)^p \\ &\le C_{2p}\max\{1,T^{p-1}\}\mathbb{E}\int_0^T \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^{2p} d\tau.\end{aligned} \tag{75}$$

Here, the latter step is again due to Jensen's inequality. The right-hand sides of (74) and (75) can be further bounded since

$$\begin{aligned}\mathbb{E}\int_0^T \left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^{2p} d\tau &\le 2^{2p-1}\mathbb{E}\int_0^T \left(\left\|V_\tau^i\right\|_2^{2p} + \left\|v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^{2p}\right)d\tau \\ &\le 2^{2p-1}\mathbb{E}\int_0^T \left(\left\|V_\tau^i\right\|_2^{2p} + 2^{p-1}\left(b_1^p + b_2^p\int \|v\|_2^{2p}\, d\widehat{\rho}_\tau^N(v)\right)\right)d\tau,\end{aligned} \tag{76}$$

where in the last step we made use of [12, Lemma 3.3], which shows that

$$\left\|v_\alpha(\widehat{\rho}_\tau^N)\right\|_2^2 \le \int \|v\|_2^2 \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\widehat{\rho}_t^N)}} d\widehat{\rho}_\tau^N(v) \le b_1 + b_2\int \|v\|_2^2\, d\widehat{\rho}_\tau^N(v),$$

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

$$b_1 = C_4^2 + b_2 \quad \text{and} \quad b_2 = 2\frac{C_2}{C_3}\left(1 + \frac{1}{\alpha C_3}\frac{1}{C_4^2}\right) \tag{77}$$

in the case that $\mathcal{E}$ satisfies the coercivity assumption (21). Inserting the upper bounds (74) and (75) together with the estimate (76) into (73) yields

$$\mathbb{E}\sup_{t\in[0,T]} \left\|V_t^i\right\|_2^{2p} \leq C\left(1 + \mathbb{E}\left\|V_0^i\right\|_2^{2p} + \mathbb{E}\int_0^T \left\|V_\tau^i\right\|_2^{2p} + \int \|v\|_2^{2p}\, d\widehat{\rho}_\tau^N(v)\, d\tau\right) \tag{78}$$

with a constant $C = C(p, \lambda, \sigma, d, T, b_1, b_2)$. Averaging (78) over i allows to bound

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,T]} \int \|v\|_2^{2p}\, d\widehat{\rho}_t^N(v) &\leq C\left(1 + \mathbb{E}\int \|v\|_2^{2p}\, d\widehat{\rho}_0^N(v) + 2\int_0^T \mathbb{E}\int \|v\|_2^{2p}\, d\widehat{\rho}_\tau^N(v)\, d\tau\right),\\ &\leq C\left(1 + \mathbb{E}\int \|v\|_2^{2p}\, d\widehat{\rho}_0^N(v) + 2\int_0^T \mathbb{E}\sup_{\widetilde{\tau}\in[0,\tau]}\int \|v\|_2^{2p}\, d\widehat{\rho}_{\widetilde{\tau}}^N(v)\, d\tau\right),\end{aligned}$$

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 $\overline{V}_t^i$ allows to derive

$$\mathbb{E}\sup_{t\in[0,T]} \left\|\overline{V}_t^i\right\|_2^{2p} \leq C\left(1 + \mathbb{E}\left\|\overline{V}_0^i\right\|_2^{2p} + \mathbb{E}\int_0^T \left\|\overline{V}_\tau^i\right\|_2^{2p} + \int \|v\|_2^{2p}\, d\rho_\tau(v)\, d\tau\right) \tag{79}$$

in place of (78). Noticing that $\int \|v\|_2^{2p}\, d\rho_\tau(v) = \mathbb{E}\left\|\overline{V}_\tau^i\right\|_2^{2p}$ 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,

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

□

Extended proof of Proposition 3.11. By exploiting the boundedness of the dynamics established in Lemma 3.10 through a cutoff technique, we can follow the steps taken in [27, 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

$$I_M(t) = \begin{cases} 1, & \text{if } \frac{1}{N}\sum_{i=1}^N \max\left\{\left\|V_\tau^i\right\|_2^4, \left\|\overline{V}_\tau^i\right\|_2^4\right\} \leq M \text{ for all } \tau \in [0,t],\\ 0, & \text{else,}\end{cases} \tag{80}$$

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 $\mathbb{E}\left\|V_t^i - \overline{V}_t^i\right\|_2^2 I_M(t)$ the inequality

$$\begin{aligned}
\mathbb{E}\left\|V_t^i - \overline{V}_t^i\right\|_2^2 I_M(t) \leq\ & 2\mathbb{E}\left\|V_0^i - \overline{V}_0^i\right\|_2^2 \\
& + 4\lambda^2\mathbb{E}\left\|\int_0^t \left(\left(V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right) - \left(\overline{V}_\tau^i - v_\alpha(\rho_\tau)\right)\right) I_M(\tau)\, d\tau\right\|_2^2 \\
& + 4\sigma^2\mathbb{E}\left\|\int_0^t \left(\left\|V_\tau^i - v_\alpha(\widehat{\rho}_\tau^N)\right\|_2 - \left\|\overline{V}_\tau^i - v_\alpha(\rho_\tau)\right\|_2\right) I_M(\tau)\, dB_\tau^i\right\|_2^2 \\
\leq\ & 2\mathbb{E}\left\|V_0^i - \overline{V}_0^i\right\|_2^2 \\
& + 8\lambda^2 T\mathbb{E}\int_0^t \left(\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 + \left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2\right) I_M(\tau)\, d\tau \\
& + 8\sigma^2\mathbb{E}\int_0^t \left(\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 + \left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2\right) I_M(\tau)\, d\tau,
\end{aligned}$$

where we used in the first step that the processes V_τ^i and $\overline{V}_\tau^i$ 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

$$\mathbb{E}\left\|V_t^i - \overline{V}_t^i\right\|_2^2 I_M(t) \leq 8\left(\lambda^2 T + \sigma^2\right)\int_0^t \mathbb{E}\left(\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 + \left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2\right) I_M(\tau)\, d\tau, \quad (81)$$

where it remains to control $\mathbb{E}\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\|_2^2 I_M(\tau)$. By means of Lemmas A.1 and A.2 below we have the bound

$$\begin{aligned}
\mathbb{E}\left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2 I_M(\tau) &\leq 2\mathbb{E}\left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\right\|_2^2 I_M(\tau) + 2\mathbb{E}\left\|v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau)\right\|_2^2 I_M(\tau) \\
&\leq C\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 I_M(\tau) + N^{-1}\right) \\
&\leq C\left(\max_{i=1,\dots,N} \mathbb{E}\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 I_M(\tau) + N^{-1}\right)
\end{aligned} \quad (82)$$

for a constant $C = C(\alpha, C_1, C_2, M, \mathcal{M}_2, b_1, b_2)$. After integrating the bound (82) into (81) and taking the maximum over i we are left with

$$\max_{i=1,\dots,N} \mathbb{E}\left\|V_t^i - \overline{V}_t^i\right\|_2^2 I_M(t) \leq C\int_0^t \max_{i=1,\dots,N} \mathbb{E}\left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 I_M(\tau)\, d\tau + CTN^{-1}, \quad (83)$$

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 $t \in [0,T]$. □

Lemma A.1. *Let I_M be as defined in (80). Then, under the assumptions of Theorem 3.2, it holds*

$$\left\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\right\|_2^2 I_M(\tau) \leq C\frac{1}{N}\sum_{i=1}^N \left\|V_\tau^i - \overline{V}_\tau^i\right\|_2^2 I_M(\tau)$$

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

Proof. The proof follows the steps taken in [12, 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

$$\begin{aligned}\frac{e^{-\alpha\underline{\mathcal{E}}}\, I_M(\tau)}{\frac{1}{N}\sum_{i=1}^N \omega_\alpha(V_\tau^i)} &\le \frac{I_M(\tau)}{\exp\big(-\alpha\frac{1}{N}\sum_{i=1}^N(\mathcal{E}(V_\tau^i)-\underline{\mathcal{E}})\big)} \le \frac{I_M(\tau)}{\exp\big(-\alpha C_2\frac{1}{N}\sum_{i=1}^N(1+\|V_\tau^i\|_2^2)\big)} \\ &\le \exp\big(\alpha C_2(1+\sqrt{M})\big) =: c_M,\end{aligned} \tag{84}$$

where, in the second inequality, we used the assumption (20) on $\mathcal{E}$. An analogous statement can be obtained for the processes $\overline{V}_\tau^i$.
For the norm of the difference between $v_\alpha(\widehat{\rho}_\tau^N)$ and $v_\alpha(\bar{\rho}_\tau^N)$ we have the decomposition

$$\begin{aligned}\big\|v_\alpha(\widehat{\rho}_\tau^N) - v_\alpha(\bar{\rho}_\tau^N)\big\|_2\, I_M(\tau) &= \left\| \frac{\sum_{i=1}^N V_\tau^i \omega_\alpha(V_\tau^i)}{\sum_{j=1}^N \omega_\alpha(V_\tau^j)} - \frac{\sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} \right\|_2 I_M(\tau) \\ &\le (\|T_1\|_2 + \|T_2\|_2 + \|T_3\|_2)\, I_M(\tau),\end{aligned} \tag{85}$$

where the terms T_1, T_2 and T_3 are obtained by inserting mixed terms with respect to V_τ^i and $\overline{V}_\tau^i$. They are defined implicitly below and their norm is bounded as follows. For the first term T_1 we have

$$\begin{aligned}\|T_1\|_2\, I_M(\tau) &= \left\| \frac{1}{N}\sum_{i=1}^N \big(V_\tau^i - \overline{V}_\tau^i\big) \frac{\omega_\alpha(V_\tau^i)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right\|_2 I_M(\tau) \\ &\le \frac{1}{N}\sum_{i=1}^N \big\|V_\tau^i - \overline{V}_\tau^i\big\|_2 \left| \frac{\omega_\alpha(V_\tau^i)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right| I_M(\tau) \\ &\le \left| \frac{e^{-\alpha\underline{\mathcal{E}}}\, I_M(\tau)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right| \frac{1}{N}\sum_{i=1}^N \big\|V_\tau^i - \overline{V}_\tau^i\big\|_2\, I_M(\tau) \\ &\le c_M \sqrt{\frac{1}{N}\sum_{i=1}^N \big\|V_\tau^i - \overline{V}_\tau^i\big\|_2^2\, I_M(\tau)},\end{aligned} \tag{86}$$

where we made use of (84) and Cauchy-Schwarz inequality in the last step. For the second term T_2, by using the assumption (19) on $\mathcal{E}$ in the third line and by following similar steps, we obtain

$$\begin{aligned}\|T_2\|_2\, I_M(\tau) &= \left\| \frac{1}{N}\sum_{i=1}^N \big(\omega_\alpha(V_\tau^i) - \omega_\alpha(\overline{V}_\tau^i)\big) \frac{\overline{V}_\tau^i}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} I_M(\tau) \right\|_2 I_M(\tau) \\ &\le \frac{1}{N}\sum_{i=1}^N \big|\omega_\alpha(V_\tau^i) - \omega_\alpha(\overline{V}_\tau^i)\big| \left\| \frac{\overline{V}_\tau^i}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right\|_2 I_M(\tau) \\ &\le \alpha C_1 e^{-\alpha\underline{\mathcal{E}}} \frac{1}{N}\sum_{i=1}^N \big(\big\|V_\tau^i\big\|_2 + \big\|\overline{V}_\tau^i\big\|_2\big) \big\|V_\tau^i - \overline{V}_\tau^i\big\|_2 \frac{\big\|\overline{V}_\tau^i\big\|_2}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} I_M(\tau) \\ &\le \frac{3}{2}\alpha C_1 \left| \frac{e^{-\alpha\underline{\mathcal{E}}} I_M(\tau)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)} \right| \sqrt{\frac{1}{N}\sum_{i=1}^N \Big(\|V_\tau^i\|_2^4 + \big\|\overline{V}_\tau^i\big\|_2^4\Big) I_M(\tau)} \\ &\qquad \cdot \sqrt{\frac{1}{N}\sum_{i=1}^N \big\|V_\tau^i - \overline{V}_\tau^i\big\|_2^2\, I_M(\tau)} \\ &\le 3\alpha C_1 c_M M^{\frac{1}{2}} \sqrt{\frac{1}{N}\sum_{i=1}^N \big\|V_\tau^i - \overline{V}_\tau^i\big\|_2^2\, I_M(\tau)}.\end{aligned} \tag{87}$$

Analogously, for the third term T_3, we get

$$\begin{aligned}
\|T_3\|_2 I_M(\tau) &= \left\| \frac{\sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(V_\tau^j)} - \frac{\sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} \right\|_2 I_M(\tau) \\
&\leq \frac{1}{N} \sum_{j=1}^N \left| \omega_\alpha(\overline{V}_\tau^j) - \omega_\alpha(V_\tau^j) \right| \left\| \frac{\frac{1}{N}\sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i)}{\left(\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)\right)\left(\frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)\right)} \right\|_2 I_M(\tau) \\
&\leq \alpha C_1 e^{-2\alpha\underline{\mathcal{E}}} \frac{1}{N} \sum_{j=1}^N \left(\left\|V_\tau^j\right\|_2 + \left\|\overline{V}_\tau^j\right\|_2 \right) \left\|V_\tau^j - \overline{V}_\tau^j\right\|_2 I_M(\tau) \\
&\qquad \cdot \frac{\frac{1}{N}\sum_{i=1}^N \left\|\overline{V}_\tau^i\right\|_2 I_M(\tau)}{\left(\frac{1}{N}\sum_{j=1}^N \omega_\alpha(V_\tau^j)\right)\left(\frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)\right)} \\
&\leq \sqrt{2}\alpha C_1 c_M^2 M^{\frac{1}{4}} \sqrt{\frac{1}{N}\sum_{j=1}^N \left(\left\|V_\tau^j\right\|_2^2 + \left\|\overline{V}_\tau^j\right\|_2^2 \right) I_M(\tau)} \sqrt{\frac{1}{N}\sum_{j=1}^N \left\|V_\tau^j - \overline{V}_\tau^j\right\|_2^2 I_M(\tau)} \\
&\leq 2\alpha C_1 c_M^2 M^{\frac{1}{2}} \sqrt{\frac{1}{N}\sum_{j=1}^N \left\|V_\tau^j - \overline{V}_\tau^j\right\|_2^2 I_M(\tau)}.
\end{aligned} \tag{88}$$

By inserting the three individual bounds (86), (87) and (88) into (85) and taking the squares of both sides, we obtain the upper bound from the statement. □

Lemma A.2. *Let $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and let I_M be as defined in (80). Then, under the assumptions of Theorem 3.2, it holds*

$$\sup_{\tau\in[0,T]} \mathbb{E} \left\| v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau) \right\|_2^2 I_M(\tau) \leq C N^{-1} \tag{89}$$

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 (77).

Proof. The proof follows the steps taken in [27, 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

$$\begin{aligned}
\left\| v_\alpha(\bar{\rho}_\tau^N) - v_\alpha(\rho_\tau) \right\|_2 I_M(\tau) &= \left\| \sum_{i=1}^N \overline{V}_\tau^i \frac{\omega_\alpha(\overline{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} - \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\rho_\tau)}} \, d\rho_\tau(v) \right\|_2 I_M(\tau) \\
&\leq \left(\|T_1\|_2 + \|T_2\|_2 \right) I_M(\tau),
\end{aligned} \tag{90}$$

where the terms T_1 and T_2 are defined implicitly and bounded in what follows. By utilizing the bound (84), for the first term T_1, we get

$$\begin{aligned}
\|T_1\|_2 I_M(\tau) &= \left\| \sum_{i=1}^N \overline{V}_\tau^i \frac{\omega_\alpha(\overline{V}_\tau^i)}{\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} - \int v \frac{\omega_\alpha(v)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} \, d\rho_\tau(v) \right\|_2 I_M(\tau) \\
&= \left| \frac{I_M(\tau)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} \right| \left\| \frac{1}{N} \sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i) - \int v\omega_\alpha(v) \, d\rho_\tau(v) \right\|_2 \\
&\leq c_M e^{\alpha\underline{\mathcal{E}}} \left\| \frac{1}{N} \sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i) - \int v\omega_\alpha(v) \, d\rho_\tau(v) \right\|_2.
\end{aligned} \tag{91}$$

Similarly, for the second term we have

$$\begin{aligned}\|T_2\|_2\, I_M(\tau) &= \left\| \int v \frac{\omega_\alpha(v)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)}\, d\rho_\tau(v) - \int v \frac{\omega_\alpha(v)}{\|\omega_\alpha\|_{L_1(\rho_\tau)}}\, d\rho_\tau(v) \right\|_2 I_M(\tau) \\ &= \left| \frac{I_M(\tau)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j)} \right| \|v_\alpha(\rho_\tau)\|_2 \left| \frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j) - \|\omega_\alpha\|_{L_1(\rho_\tau)} \right| \\ &\leq c_M e^{\alpha \underline{\mathcal{E}}} \sqrt{b_1 + b_2 \mathcal{M}_2} \left| \frac{1}{N}\sum_{j=1}^N \omega_\alpha(\overline{V}_\tau^j) - \int \omega_\alpha(v)\, d\rho_\tau(v) \right|, \end{aligned} \tag{92}$$

where the last step uses that by Jensen's inequality and [12, 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 (77) 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 3.2 as a consequence of the initial regularity $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$. In order to further bound (91) and (92), respectively, let us introduce the random variables

$$Z_\tau^i = \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i) - \int v \omega_\alpha(v)\, d\rho_\tau(v) \quad \text{and} \quad z_\tau^i = \omega_\alpha(\overline{V}_\tau^i) - \int \omega_\alpha(v)\, d\rho_\tau(v),$$

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

$$\frac{1}{N}\sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i) - \int v \omega_\alpha(v)\, d\rho_\tau(v) = \frac{1}{N}\sum_{i=1}^N Z_\tau^i$$

and

$$\frac{1}{N}\sum_{i=1}^N \omega_\alpha(\overline{V}_\tau^j) - \int \omega_\alpha(v)\, d\rho_\tau(v) = \frac{1}{N}\sum_{i=1}^N z_\tau^i,$$

respectively. Moreover, due to the independence of the $\overline{V}_\tau^i$'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

$$\begin{aligned}\mathbb{E} \left\| \frac{1}{N}\sum_{i=1}^N \overline{V}_\tau^i \omega_\alpha(\overline{V}_\tau^i) - \int v \omega_\alpha(v)\, d\rho_\tau(v) \right\|_2^2 &= \mathbb{E} \left\| \frac{1}{N}\sum_{i=1}^N Z_\tau^i \right\|_2^2 = \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \sum_{j=1}^N \langle Z_\tau^i, Z_\tau^j \rangle \\ &= \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \left\| Z_\tau^i \right\|_2^2 = \frac{1}{N} \mathbb{E} \left\| Z_\tau^1 \right\|_2^2 \leq 4 e^{-\alpha \underline{\mathcal{E}}} \mathcal{M}_2 \frac{1}{N}, \end{aligned} \tag{93}$$

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

$$\begin{aligned}\mathbb{E} \left\| Z_\tau^1 \right\|_2^2 &\leq 2 \mathbb{E} \left\| \overline{V}_\tau^1 \omega_\alpha(\overline{V}_\tau^1) \right\|_2^2 + 2 \left\| \int v \omega_\alpha(v)\, d\rho_\tau(v) \right\|_2^2 \\ &\leq 2 e^{-\alpha \underline{\mathcal{E}}} \left(\mathbb{E} \left\| \overline{V}_\tau^1 \right\|_2^2 + \int \|v\|_2^2\, d\rho_\tau(v) \right) \leq 4 e^{-\alpha \underline{\mathcal{E}}} \mathcal{M}_2. \end{aligned}$$

Following analogous arguments and noting that

$$\mathbb{E} \left| z_\tau^1 \right|^2 \leq 2 \mathbb{E} \left| \omega_\alpha(\overline{V}_\tau^j) \right|^2 + 2 \left| \int \omega_\alpha(v)\, d\rho_\tau(v) \right|^2 \leq 4 e^{-\alpha \underline{\mathcal{E}}}$$

yields the inequality

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}\omega_\alpha(\overline{V}_\tau^j)-\int\omega_\alpha(v)\,d\rho_\tau(v)\right|^2=\frac{1}{N}\mathbb{E}\left|z_\tau^1\right|^2\leq 4e^{-\alpha\underline{\mathcal{E}}}\frac{1}{N}. \tag{94}$$

Taking the square and expectation on both sides of (91) and (92), inserting the two individual bounds (93) and (94), gives the statement after recalling (90). □