CAMBRIDGE
UNIVERSITY PRESS

PAPER

Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law

Konstantin Riedl[1,2]

[1]Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, Munich, Germany and [2]Munich Center for Machine Learning, Munich, Germany
Email: konstantin.riedl@ma.tum.de

Abstract

In this paper, we study consensus-based optimisation (CBO), a versatile, flexible and customisable optimisation method suitable for performing nonconvex and nonsmooth global optimisations in high dimensions. CBO is a multi-particle metaheuristic, which is effective in various applications and at the same time amenable to theoretical analysis thanks to its minimalistic design. The underlying dynamics, however, is flexible enough to incorporate different mechanisms widely used in evolutionary computation and machine learning, as we show by analysing a variant of CBO which makes use of memory effects and gradient information. We rigorously prove that this dynamics converges to a global minimiser of the objective function in mean-field law for a vast class of functions under minimal assumptions on the initialisation of the method. The proof in particular reveals how to leverage further, in some applications advantageous, forces in the dynamics without loosing provable global convergence. To demonstrate the benefit of the herein investigated memory effects and gradient information in certain applications, we present numerical evidence for the superiority of this CBO variant in applications such as machine learning and compressed sensing, which en passant widen the scope of applications of CBO.

1. Introduction

Interacting multi-particle systems are ubiquitous in a wide variety of scientific disciplines with application areas reaching from atomic scales over the human scale to the astronomical scale. For instance, large-scale multi-agent models are used to understand the coordinated movement of animal groups [19, 51] or crowds of people [1, 20]. Especially fascinating in this context is that such complex and often intelligent behaviour – phenomena known as self-organisation and swarm intelligence – emerge from seemingly simple rules of interaction [56]. This intriguing capabilities have drawn researchers' attention towards designing interacting particle systems for specific purposes in various disciplines. In applied mathematics in particular, agent-based optimisation algorithms look back on a long and successful history of empirically achieving state-of-the-art performance on challenging global unconstrained problems of the form

$$x^* = \underset{x \in \mathbb{R}^d}{\arg \min} \, \mathcal{E}(x).$$

Here, $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ denotes a possibly nonconvex and nonsmooth high-dimensional objective function, whose global minimiser x^* is assumed to exist and be unique for the remainder of this work. Well-known representatives of this family are Evolutionary Programming [24], Genetic Algorithms [38], Particle

Swarm Optimisation [43] and Ant Colony Optimisation [23]. They belong to the broad class of so-called metaheuristics [4, 6], which are methods orchestrating an interaction between local improvement procedures and global strategies, deterministic and stochastic processes, to eventually design an efficient and robust procedure for searching the solution space of the objective function $\mathcal{E}$.

Motivated by both the substantiated success of metaheuristics in applications and the lack of rigorous theoretical guarantees about their convergence and performance, the authors of [52] proposed consensus-based optimisation (CBO), which follows the spirit of metaheuristics but allows for a rigorous theoretical analysis [12, 14, 28, 29, 33, 34]. By taking inspiration from consensus formation in opinion dynamics [36], CBO methods use N particles $X^1, \ldots, X^N$ to explore the energy landscape of the objective $\mathcal{E}$ and to eventually form a consensus about the location of the global minimiser x^*. In its original form [52], the dynamics of each particle X^i, which is governed by a stochastic differential equation (SDE), is subject to two competing forces. A deterministic drift term pulls the particles towards a so-called consensus point, which is an instantaneously computed weighted average of the positions of all particles and approximates the global minimiser x^* the best possible given the currently available information. The resulting contractive behaviour is counteracted by the second term which is stochastic in nature and thereby features the exploration of the energy landscape of the objective function. Its magnitude and therefore its explorative power scales with the distance of the individual particle from the consensus point, which encourages particles far away to explore larger regions of the domain, while particles already close advance their position only locally.

In this work, motivated by the numerical evidence presented below as well as other recent papers such as [32, 54, 55], we consider a more elaborate variant of this dynamics which exhibits the two following additional drift terms.

- The first is a drift towards the historical best position of the particular particle. To store such information, we follow the work [32], where the authors introduce for each particle an additional state variable Y^i, which can be regarded as the memory of the respective particle X^i. In contrast to the original dynamics, an individual particle is therefore described by the tuple (X^i, Y^i). Moreover, the consensus point is no longer computed from the instantaneous positions X^i, but the historical best positions Y^i.

- The second term is a drift in the direction of the negative gradient of $\mathcal{E}$ evaluated at the current position of the respective particle X^i.

Both terms are accompanied by associated noise terms. We now make the CBO dynamics with memory effects and gradient information rigorous by providing a formal description of the interacting particle system. A visualisation of the dynamics with all relevant quantities and forces is provided in Figure 1. Given a finite time horizon $T > 0$, and user-specified parameters $\alpha, \beta, \theta, \kappa, \lambda_1, \sigma_1 > 0$ and $\lambda_2, \lambda_3, \sigma_2, \sigma_3 \geq 0$, the dynamics is given by the system of SDEs

$$
\begin{aligned}
dX_t^i = &-\lambda_1\big(X_t^i - y_\alpha(\widehat{\rho}_{Y,t}^N)\big)\, dt - \lambda_2\big(X_t^i - Y_t^i\big)\, dt - \lambda_3 \nabla\mathcal{E}(X_t^i)\, dt \\
&+ \sigma_1 D\big(X_t^i - y_\alpha(\widehat{\rho}_{Y,t}^N)\big)\, dB_t^{1,i} + \sigma_2 D\big(X_t^i - Y_t^i\big)\, dB_t^{2,i} + \sigma_3 D\big(\nabla\mathcal{E}(X_t^i)\big)\, dB_t^{3,i},
\end{aligned}
\tag{1.1a}
$$

$$
dY_t^i = \kappa\,\big(X_t^i - Y_t^i\big)\, S^{\beta,\theta}\big(X_t^i, Y_t^i\big)\, dt
\tag{1.1b}
$$

for $i = 1, \ldots, N$ and where $((B_t^{m,i})_{t\geq 0})_{i=1,\ldots,N}$ are independent standard Brownian motions in $\mathbb{R}^d$ for $m \in \{1, 2, 3\}$. The system is complemented with independent initial data $(X_0^i, Y_0^i)_{i=1,\ldots,N}$, typically such that $X_0^i = Y_0^i$ for all $i = 1, \ldots, N$. A numerical implementation of the scheme usually originates from an Euler-Maruyama time discretisation of equation (1.1). The first term appearing in the SDE for the position X_t^i, i.e., in the first line of equation (1.1a), is the drift towards the consensus point

$$
y_\alpha(\widehat{\rho}_{Y,t}^N) := \int y\, \frac{\omega_\alpha(y)}{\|\omega_\alpha\|_{L_1(\widehat{\rho}_{Y,t}^N)}}\, d\widehat{\rho}_{Y,t}^N(y), \quad \text{with} \quad \omega_\alpha(y) := \exp\big(-\alpha\mathcal{E}(y)\big).
\tag{1.2}
$$

Here, $\widehat{\rho}_{Y,t}^N$ denotes the random empirical measure of the particles' historical best positions, i.e., $\widehat{\rho}_{Y,t}^N := \frac{1}{N}\sum_{i=1}^N \delta_{Y_t^i}$. Definition (1.2) is motivated by the fact that $y_\alpha(\widehat{\rho}_{Y,t}^N) \approx \arg\min_{i\in\{1,\ldots,N\}} \mathcal{E}(Y_t^i)$ as $\alpha \to \infty$ under
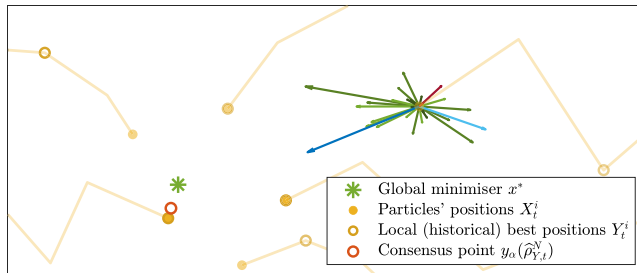
Figure 1. *A visualisation of the CBO dynamics (1.1) with memory effects and gradient information. Particles with positions $X^1, \ldots, X^N$ (yellow dots with their trajectories) explore the energy landscape of the objective $\mathcal{E}$ in search of the global minimiser x^* (green star). Each particle stores its local historical best position Y_t^i (yellow circles). The dynamics of the position X_t^i of each particle is governed by three deterministic terms with associated random noise terms (visualised by depicting eight possible realisations with differently shaded green arrows). A global drift term (dark blue arrow) drags the particle towards the consensus point $y_\alpha(\widehat{\rho}_{Y,t}^N)$ (orange circle), which is computed as a weighted (visualised through colour opacity) average of the particles' historical best positions. A local drift term (light blue arrow) imposes movement towards the respective local best position Y_t^i. A gradient drift term (purple arrow) exerts a force in the direction $-\nabla \mathcal{E}(X_t^i)$.*

reasonable assumptions. The first term in the second line of equation (1.1a) is with the consensus drift associated diffusion term, which injects randomness into the dynamics and thereby features the explorative nature of the algorithm. The two commonly studied diffusion types are isotropic [12, 28, 52] and anisotropic [14, 29] diffusion with

$$
D(\cdot) = \begin{cases} \|\cdot\|_2 \,\mathrm{Id}, & \text{for isotropic diffusion,} \\ \mathrm{diag}\,(\cdot), & \text{for anisotropic diffusion,} \end{cases} \tag{1.3}
$$

where $\mathrm{Id} \in \mathbb{R}^{d \times d}$ is the identity matrix and $\mathrm{diag} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ the operator mapping a vector onto a diagonal matrix with the vector as its diagonal. Despite the potential of the dynamics getting trapped in affine subspaces, the coordinate-dependent scaling of anisotropic diffusion has proven to be beneficial for the performance of the method in high-dimensional applications by allowing for dimension-independent convergence rates [14, 29]. For this reason, we restrict our attention to the case of anisotropic noise in what follows. Nevertheless, theoretically similar results as the ones presented in this work can be obtained also for the isotropic case. The second term in the first line of equation (1.1a) is the drift towards the historical best position of the respective particle. In contrast to the global nature of the consensus drift, which incorporates information from all N particles, this term depends only on the past of the specific particle. To store such information about the history of each particle [32], an additional state variable Y^i is introduced for every particle, which evolves according to equation (1.1b), where

$$
S^{\beta, \theta}(x, y) = \frac{1}{2}\big(1 + \theta + \tanh\big(\beta\,(\mathcal{E}(y) - \mathcal{E}(x))\big)\big) \tag{1.4}
$$

is chosen throughout this article, which is an approximation to the Heaviside function $H(x, y) = \mathbb{1}_{\mathcal{E}(x) < \mathcal{E}(y)}$ as $\theta \to 0$ and $\beta \to \infty$. The variable Y_t^i can therefore be regarded as the memory of the ith particle, i.e., as the location of the in-time best-seen position of X^i up to time t. This can be understood most easily when discretising (1.1b) as

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

and noting that with parameter choices $\kappa = 1/\Delta t$, $\theta = 0$ and $\beta \gg 1$ it holds $Y_{k+1}^i = X_{k+1}^i$ if $\mathcal{E}(X_{k+1}^i) < \mathcal{E}(Y_k^i)$ and $Y_{k+1}^i = Y_k^i$ else. The third term in the first line of equation (1.1a) is the drift in the direction

of the negative gradient of $\mathcal{E}$, which is a local and instantaneous contribution. The remaining two terms are noise terms, which are associated with the formerly described memory and gradient drifts.

A theoretical convergence analysis of CBO can be carried out either by directly investigating the microscopic system (1.1) or its numerical time discretisation, as promoted for instance in a simplified setting in the works [33, 34], or alternatively, as done for example in [12, 14, 26, 28, 29], by analysing the macroscopic behaviour of the particle density through a mean-field limit associated with (1.1). Formally, such mean-field limit is given by the self-consistent nonlinear and nonlocal SDE

$$
\begin{aligned}
d\overline{X}_t = &-\lambda_1\big(\overline{X}_t - y_\alpha(\rho_{Y,t})\big)\,dt - \lambda_2\big(\overline{X}_t - \overline{Y}_t\big)\,dt - \lambda_3\nabla\mathcal{E}(\overline{X}_t)\,dt \\
&+ \sigma_1 D\big(\overline{X}_t - y_\alpha(\rho_{Y,t})\big)\,dB_t^1 + \sigma_2 D\big(\overline{X}_t - \overline{Y}_t\big)\,dB_t^2 + \sigma_3 D\big(\nabla\mathcal{E}(\overline{X}_t)\big)\,dB_t^3,
\end{aligned}
\tag{1.5a}
$$
$$
d\overline{Y}_t = \kappa\,\big(\overline{X}_t - \overline{Y}_t\big)\,S^{\beta,\theta}\big(\overline{X}_t, \overline{Y}_t\big)\,dt,
\tag{1.5b}
$$

which is complemented with initial datum $(\overline{X}_0, \overline{Y}_0) \sim \rho_0$, and where $\rho_t = \rho(t) = \mathrm{Law}\,\big(\overline{X}_t, \overline{Y}_t\big)$ with marginal law $\rho_{Y,t}$ of $\overline{Y}_t$ given by $\rho_{Y,t} = \rho_Y(t, \,\cdot\,) = \int d\rho_t(\,\cdot\,, y)$. The measure $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ in particular weakly satisfies the Fokker-Planck equation

$$
\begin{aligned}
\partial_t\rho_t = &\,\mathrm{div}_x\big(\big(\lambda_1\big(x - y_\alpha(\rho_{Y,t})\big) + \lambda_2(x - y) + \lambda_3\nabla\mathcal{E}(x)\big)\rho_t\big) + \mathrm{div}_y\big(\big(\kappa(y - x)S^{\beta,\theta}(x, y)\big)\rho_t\big) \\
&+ \frac{1}{2}\sum_{k=1}^{d}\partial_{x_k x_k}^2\left(\Big(\sigma_1^2 D\big(x - y_\alpha(\rho_{Y,t})\big)_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla\mathcal{E}(x))_{kk}^2\Big)\rho_t\right),
\end{aligned}
\tag{1.6}
$$

see Definition 2.1. Working with the partial differential equation (PDE) (1.6) instead of the interacting particle system (1.1) typically permits to employ more powerful technical tools, which result in stronger and deterministic statements about the long-time behaviour of the average agent density ρ. This analysis approach is rigorously justified by the mean-field approximation, i.e., the fact that the empirical particle measure $\widehat{\rho}_t^N := \frac{1}{N}\sum_{i=1}^{N}\delta_{(X_t^i, Y_t^i)}$ converges in some sense to the mean-field law ρ_t as the number of particles N tends to infinity. For the original CBO dynamics, a qualitative result about convergence in distribution is provided in [39], which is based on a tightness argument in the path space. More precisely, the authors of that work show that the sequence $\{\widehat{\rho}^N\}_{N\geq 2}$ of $\mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$-valued random variables is tight, which permits to employ Prokhorov's theorem to obtain, up to a subsequence, some limiting measure, which turns out to be deterministic and at the same time satisfy the associated mean-field PDE. A more desirable quantitative approximation result, on the other hand, can be established by proving propagation of chaos, i.e., by establishing for instance

$$
\max_{i=1,\dots,N}\sup_{t\in[0,T]}\mathbb{E}\left\|(X_t^i, Y_t^i) - (\overline{X}_t^i, \overline{Y}_t^i)\right\|_2^2 \leq CN^{-1} \qquad \text{as } N \to \infty,
$$

where $(\overline{X}_t^i, \overline{Y}_t^i)$ denote N i.i.d. copies of the mean-field dynamics (1.5). For the original variant of unconstrained CBO, this was first done in [28, Section 3.3]. To keep the focus of this work on the long-time behaviour of the CBO variant (1.6), a rigorous analysis of the mean-field approximation is left for future considerations.

Before summarising the contributions of the present paper, let us put our work into context by providing a comprehensive literature overview about the history, developments and achievements of CBO.

Versatility and flexibility of CBO: a literature overview

Since its introduction in the work [52], CBO has gained a significant amount of attention from various research groups. This has led to a vast variety of different developments, of both theoretical and applied nature, as well as what concerns the mathematical modelling and numerical analysis of the method. By interpreting CBO as a stochastic relaxation of gradient descent, the recent work [53] even establishes a connection between the worlds of derivative-free and gradient-based optimisation.

A first rigorous but local convergence proof of the mean-field limit of CBO to global minimisers is provided for the cases of isotropic and anisotropic diffusion in [12, 14], respectively. By analysing

the time-evolution of the variance of the law of the mean-field dynamics ρ_t and proving its exponential decay towards zero, the authors first establish consensus formation at some stationary point before they ensure that this consensus is actually close to the global minimiser. A similarly flavoured approach is pursued in [33, 34], however, directly for the fully in-time discrete microscopic system and in the simplified setting, where the same Brownian motion is used for all agents, which limits the exploration capabilities of the method. In contrast, in the recent works [28, 29], the authors, again for the isotropic and anisotropic CBO variant, respectively, investigate the time-evolution of the Wasserstein-2 distance between the law ρ_t and a Dirac delta at the global minimiser. This is also the strategy which we pursue in this paper. By proving the exponential decay of $W_2(\rho_t, \delta_{x^*})$ to zero, consensus at the desired location follows immediately. Moreover, by providing a probabilistic quantitative result about the mean-field approximation, the authors give a first, and so far unique, holistic global convergence proof for the implementable, i.e., discretised numerical CBO algorithm in the unconstrained case. The results about the mean-field approximation of the latter papers were partially inspired by the series of works [25–27], in which the authors constrain the particle dynamics of CBO to compact hypersurfaces and prove local convergence of the numerical scheme to minimisers by adapting the technique of [12, 14]. This ensures a beneficial compactness of the stochastic processes, which simplifies the convergence of the interacting particle dynamics to the mean-field dynamics. In the unconstrained case, such intrinsic compactness is replaced by the fact that the dynamics are bounded with high probability, which is sufficient to establish convergence in probability. Further related works about CBO for optimisations with constraints include the papers [35, 44], where a problem on the Stiefel manifold is approached, and the works [9, 15], where the constrained optimisation is recast into a penalised problem. The philosophy of using an interacting swarm of particles to approach various relevant problems in science and engineering has promoted several variations of the original CBO algorithm for minimisation. Amongst them are methods based on consensus dynamics to tackle multi-objective optimisation problems [7, 8, 45], saddle point problems [40], the search for several minimisers simultaneously [10] or the sampling from certain distributions [13].

In the same vein and also in the spirit of this work, the original CBO method itself has undergone several modifications allowing for a more complex dynamics. This includes the use of particles with memory [31, 55], the integration of momentum [17], the usage of jump-diffusion processes [42] and the exploitation of on-the-fly extracted higher-order differential information through inferred gradients based on point evaluations of the objective function [54]. It moreover turned out that the renowned particle swarm optimisation method (PSO) [43] can be formulated and regarded as a second-order generalisation of CBO [18, 32]. This insight has enabled to adapt the for CBO-developed analysis techniques to rigorously prove the convergence of PSO [41].

In the collection of formerly referenced works and beyond, CBO has demonstrated to be a valuable method for a wide scope of applications reaching from the phase retrieval or robust subspace detection problem in signal processing [26, 27], over the training of neural networks for image classification in machine learning [14, 29] as well as in the setting of clustered federated learning [16], to asset allocation in finance [5]. It has been furthermore employed to approximate low-frequency functions in the presence of high-frequency noise and to the task of solving PDEs with low-regularity solutions [17].

Contributions

In view of the various developments and the wide scope of applications, a theoretical understanding of the long-time behaviour of the in practical applications employed CBO methods is of paramount interest. In this work, we analyse a variant of CBO which incorporates memory effects as well as gradient information from a theoretical and numerical perspective. As demonstrated concisely in Figure 2 and more comprehensively in Section 4, the herein investigated dynamics, which is more involved than standard CBO, proves to be beneficial in applications in machine learning and compressed sensing. Despite this additional complexity, by employing the analysis technique devised in [28, 29], we are able to provide
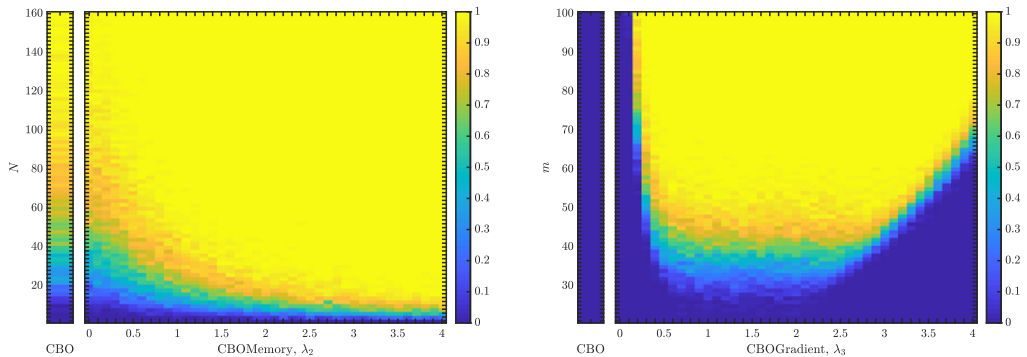
(a) *Memory effects and an additional drift towards the historical best position of each individual particle improve the success probability of CBO.*

(b) *Gradient information and a drift in the direction of the negative gradient can be indispensable in certain applications such as compressed sensing.*

Figure 2. *A demonstration of the benefits of memory effects and gradient information in CBO methods. In both settings (a) and (b) the depicted success probabilities are averaged over 100 runs of CBO and the implemented scheme is given by a Euler-Maruyama discretisation of equation (1.1) with time horizon $T = 20$, discrete time step size $\Delta t = 0.01$, $\alpha = 100$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, $\lambda_1 = 1$ and $\sigma_1 = \sqrt{1.6}$. In (a) we plot the success probability of CBO without (left separate column) and with (right phase diagram) memory effects for different values of the parameter λ_2, i.e., for different strengths of the memory drift, when optimising the Rastrigin function $\mathcal{E}(x) = \sum_{k=1}^{d} x_k^2 + \frac{5}{2}(1 - \cos(2\pi x_k))$ in dimension $d = 4$. As remaining parameters we choose $\sigma_2 = \lambda_1 \sigma_1$ and $\lambda_3 = \sigma_3 = 0$, i.e., no gradient information is involved. We observe that an increasing amount of memory drift improves the success probability significantly, even in the case where, theoretically, there are no convergence guarantees anymore, see Theorem 2.5 and Corollary 2.6. Section 4.2 provides further details. In (b) we depict the success probability of CBO without (left separate column) and with (right phase diagram) gradient information for different values of the parameter λ_3, i.e., for different strengths of the gradient drift, when solving a compressed sensing problem in dimension $d = 200$ with sparsity $s = 8$. On the vertical axis we depict the number of measurements m, from which we try to recover the sparse signal by solving the associated ℓ_1-regularised problem (LASSO). As remaining parameters we use merely $N = 10$ particles, choose $\sigma_3 = 0$ and $\lambda_2 = \sigma_2 = 0$, i.e., no memory drift is involved. We observe that gradient information is required to be able to identify the correct sparse solution and standard CBO would fail in such task. Section 4.4 provides more details.*

rigorous mean-field-type convergence guarantees to the global minimiser, which describe the behaviour of the method in the large-particle limit and allow to draw conclusions about the typically observed performance in the practicable regime. Our results for CBO with memory effects and gradient information hold for a vast class of objective functions under minimal assumptions on the initialisation of the method. Moreover, the proof reveals how to leverage further, in other applications advantageous, forces in the dynamics while still being amenable to theory and allowing for provable global convergence.

1.1. Organisation

In Section 2, after providing details about the existence of solutions to the macroscopic SDE (1.5) and the associated PDE (1.6), we present and discuss our main theoretical contribution. It is about the convergence of CBO with memory effects and gradient information, as given in equation (1.1), to the global minimiser of the objective function in mean-field law, see [28, Definition 1]. More precisely, we show that the mean-field dynamics (1.5) and (1.6) converge with exponential rate to the global minimiser.

Section 3 contains the proof details of this result. In Section 4, we numerically demonstrate the benefits of the additional memory effects and gradient information of the previously analysed CBO variant. We in particular present applications of CBO in machine learning and compressed sensing, before we conclude the paper in Section 5.

For the sake of reproducible research, in the GitHub repository https://github.com/ KonstantinRiedl/CBOGlobalConvergenceAnalysis we provide the Matlab code implementing the CBO algorithm with memory effects and gradient information analysed in this work.

1.2. Notation

Given a set $A \subset \mathbb{R}^d$, we write $(A)^c$ to denote its complement, i.e., $(A)^c := \{z \in \mathbb{R}^d : z \notin A\}$. For ℓ_∞ balls in $\mathbb{R}^d$ with centre z and radius r, we write $B_r^\infty(z)$. The space of continuous functions $f : X \to Y$ is denoted by $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$ and a suitable topological space Y. For $X \subset \mathbb{R}^n$ open and for $Y = \mathbb{R}^m$, the function space $\mathcal{C}_c^k(X, Y)$ contains functions $f \in \mathcal{C}(X, Y)$ that are k-times continuously differentiable and have compact support. Y is omitted in the case of real-valued functions. The operator ∇ denotes the standard gradient of a function on $\mathbb{R}^d$.

In this paper, we mostly study laws of stochastic processes, $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$, and we refer to a snapshot of such law at time t by writing $\rho_t \in \mathcal{P}(\mathbb{R}^d)$. Here, $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all Borel probability measures ϱ over $\mathbb{R}^d$. In $\mathcal{P}_p(\mathbb{R}^d)$ we moreover collect measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite pth moment. For any $1 \le p < \infty$, W_p denotes the Wasserstein-p distance between two Borel probability measures $\varrho_1, \varrho_2 \in \mathcal{P}_p(\mathbb{R}^d)$, see, e.g., [2]. $\mathbb{E}(\varrho)$ denotes the expectation of a probability measure ϱ.

2. Global convergence in mean-field law

In the first part of this section, we provide an existence result about solutions of the nonlinear macroscopic SDE (1.5), respectively, the associated Fokker-Planck equation (1.6). Thereafter we specify the class of studied objective functions and present the main theoretical result about the convergence of the dynamics (1.5) and (1.6) to the global minimiser.

Throughout this work, we consider the – in typical applications beneficial – case of CBO with anisotropic diffusion, i.e., $D(\cdot) = \mathrm{diag}\,(\cdot)$ in equations (1.1), (1.5) and (1.6), and also equation (2.1) below. However, up to minor modifications, analogous results can be obtained for isotropic diffusion.

2.1. Definition and existence of weak solutions

Let us begin by rigorously defining weak solutions of the Fokker-Planck equation (1.6).

Definition 2.1. *Let $\rho_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, $T > 0$. We say $\rho \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ satisfies the Fokker-Planck equation (1.6) with initial condition ρ_0 in the weak sense in the time interval $[0, T]$, if we have for all $\phi \in \mathcal{C}_c^2(\mathbb{R}^d \times \mathbb{R}^d)$ and all $t \in (0, T)$*

$$
\begin{aligned}
\frac{d}{dt} \iint \phi(x, y)\, d\rho_t(x, y) = &-\iint \kappa S^{\beta,\theta}(x, y) \left\langle y - x, \nabla_y \phi(x, y) \right\rangle d\rho_t(x, y) \\
&- \iint \lambda_1 \langle x - y_\alpha(\rho_{Y,t}), \nabla_x \phi(x, y) \rangle + \lambda_2 \langle x - y, \nabla_x \phi(x, y) \rangle + \lambda_3 \langle \nabla \mathcal{E}(x), \nabla_x \phi(x, y) \rangle\, d\rho_t(x, y) \\
&+ \frac{1}{2} \iint \sum_{k=1}^d \left(\sigma_1^2 D\big(x - y_\alpha(\rho_{Y,t})\big)_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla \mathcal{E}(x))_{kk}^2 \right) \partial_{x_k x_k}^2 \phi(x, y)\, d\rho_t(x, y)
\end{aligned}
\tag{2.1}
$$

and $\lim_{t \to 0} \rho_t = \rho_0$ (in the sense of weak convergence of measures).

For solutions of the mean-field dynamics (1.5) and (1.6), we have the following existence result.

Theorem 2.2. *Let $T > 0$, $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$. Let $\mathcal{E}: \mathbb{R}^d \to \mathbb{R}$ with $\underline{\mathcal{E}} > -\infty$ satisfy for some constants $C_1, C_2 > 0$ the conditions*

$$|\mathcal{E}(x) - \mathcal{E}(x')| \leq C_1 \left(\|x\|_2 + \|x'\|_2\right) \|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d, \tag{2.2}$$

$$\mathcal{E}(x) - \underline{\mathcal{E}} \leq C_2 \left(1 + \|x\|_2^2\right), \quad \text{for all } x \in \mathbb{R}^d, \tag{2.3}$$

and either $\sup_{x \in \mathbb{R}^d} \mathcal{E}(x) < \infty$ or

$$\mathcal{E}(x) - \underline{\mathcal{E}} \geq C_3 \|x\|_2^2, \quad \text{for all } \|x\|_2 \geq C_4 \tag{2.4}$$

for some $C_3, C_4 > 0$. Furthermore, in the case of an active gradient drift in the CBO dynamcis (1.5), i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ and obey additionally

$$\|\nabla \mathcal{E}(x) - \nabla \mathcal{E}(x')\|_2 \leq \widetilde{L}_{\nabla \mathcal{E}} \|x - x'\|_2, \quad \text{for all } x, x' \in \mathbb{R}^d \tag{2.5}$$

for some $\widetilde{L}_{\nabla \mathcal{E}} > 0$. Then, if $(\overline{X}_0, \overline{Y}_0)$ is distributed according to ρ_0, there exists a nonlinear process $(\overline{X}, \overline{Y}) \in \mathcal{C}([0, T], \mathbb{R}^d \times \mathbb{R}^d)$ satisfying (1.5) with associated law $\rho = \text{Law}((\overline{X}, \overline{Y}))$ having regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ and being a weak solution to the Fokker-Planck equation (1.6) with $\rho(0) = \rho_0$.

Assumption (2.2) requires that $\mathcal{E}$ is locally Lipschitz continuous with the Lipschitz constant being allowed to have linear growth. This entails in particular that the objective has at most quadratic growth at infinity as formulated explicitly in Assumption (2.3), which can be seen when choosing $x' = x^*$ and $C_2 = 2C_1 \max\{1, \|x^*\|_2^2\}$ in (2.2). Assumption (2.4), on the other hand, assumes that $\mathcal{E}$ also has at least quadratic growth in the farfield, i.e., overall it grows quadratically far away from x^*. Alternatively, $\mathcal{E}$ may be bounded from above. Since the objective $\mathcal{E}$ can be usually modified for the purpose of analysis outside a sufficiently large region, these growth conditions are not really restrictive. In case of an additional gradient drift term in the dynamics, i.e., $\lambda_3 \neq 0$, the objective naturally needs to be continuously differentiable. Furthermore, Assumption (2.5) imposes $\mathcal{E}$ to be $\widetilde{L}_{\nabla \mathcal{E}}$-smooth, i.e., having an $\widetilde{L}_{\nabla \mathcal{E}}$-Lipschitz continuous gradient.

Remark 2.3. *The regularity $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ obtained in Theorem 2.2 above is an immediate consequence of the regularity of the initial condition $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$. It allows to extend the test function space $\mathcal{C}_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ in Definition 2.1 to the larger space*

$$\mathcal{C}_*^2(\mathbb{R}^d \times \mathbb{R}^d) := \left\{ \phi \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d): \left|\partial_{x_k}\phi(x, y)\right| \leq C_\phi(1 + \|x\|_2 + \|y\|_2) \text{ for some } C_\phi > 0 \right.$$
$$\left. \text{and} \sup_{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d} \max_{k=1,\dots,d} |\partial_{x_k x_k}^2 \phi(x, y)| < \infty \right\}, \tag{2.6}$$

as can be seen from the proof of Theorem 2.2, which we sketch in what follows.

Proof sketch of Theorem 2.2. The proof is based on the Leray-Schauder fixed point theorem and follows the steps taken in [12, Theorems 3.1, 3.2].

Step 1: For a given function $u \in \mathcal{C}([0, T], \mathbb{R}^d)$ and an initial measure $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$, according to standard SDE theory [3, Chapters 6], we can uniquely solve the auxiliary SDE

$$d\widetilde{X}_t = -\lambda_1(\widetilde{X}_t - u_t)\, dt - \lambda_2(\widetilde{X}_t - \widetilde{Y}_t)\, dt - \lambda_3 \nabla \mathcal{E}(\widetilde{X}_t)\, dt$$
$$+ \sigma_1 D(\widetilde{X}_t - u_t)\, dB_t^1 + \sigma_2 D(\widetilde{X}_t - \widetilde{Y}_t)\, dB_t^2 + \sigma_3 D(\nabla \mathcal{E}(\widetilde{X}_t))\, dB_t^3$$
$$d\widetilde{Y}_t = \kappa(\widetilde{X}_t - \widetilde{Y}_t)\, S^{\beta,\theta}(\widetilde{X}_t, \widetilde{Y}_t)\, dt$$

with $(\widetilde{X}_0, \widetilde{Y}_0) \sim \rho_0$. This is due to the fact that the coefficients of the drift and diffusion terms are locally Lipschitz continuous and have at most linear growth, which, in turn, is a consequence of the assumptions on $\mathcal{E}$ as well as the smoothness of $S^{\beta,\theta}$ as defined in (1.4). This induces $\widetilde{\rho}_t = \text{Law}((\widetilde{X}_t, \widetilde{Y}_t))$. Moreover, the assumed regularity of the initial distribution $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$ allows to obtain a fourth-order moment estimate of the form $\mathbb{E}[\|\widetilde{X}_t\|_2^4 + \|\widetilde{Y}_t\|_2^4] \leq (1 + 2\mathbb{E}[\|\widetilde{X}_0\|_2^4 + \|\widetilde{Y}_0\|_2^4])e^{ct}$, see, e.g. [3, Chapter 7]. So, in particular, $\widetilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$.

Step 2: For some test function $\phi \in \mathcal{C}_*^2(\mathbb{R}^d \times \mathbb{R}^d)$ as defined in (2.6), by Itô's formula, we derive

$$d\phi(\widetilde{X}_t, \widetilde{Y}_t) = \nabla_x \phi(\widetilde{X}_t, \widetilde{Y}_t) \cdot \left(-\lambda_1(\widetilde{X}_t - u_t) - \lambda_2(\widetilde{X}_t - \widetilde{Y}_t) - \lambda_3 \nabla\mathcal{E}(\widetilde{X}_t) \right) dt$$
$$+ \nabla_y \phi(\widetilde{X}_t, \widetilde{Y}_t) \cdot \left(\kappa(\widetilde{X}_t - \widetilde{Y}_t) S^{\beta,\theta}(\widetilde{X}_t, \widetilde{Y}_t) \right) dt$$
$$+ \frac{1}{2} \sum_{k=1}^d \partial^2_{x_k x_k} \phi(\widetilde{X}_t, \widetilde{Y}_t) \left(\sigma_1^2 D(\widetilde{X}_t - u_t)_{kk}^2 + \sigma_2^2 D(\widetilde{X}_t - \widetilde{Y}_t)_{kk}^2 + \sigma_3^2 D(\nabla\mathcal{E}(\widetilde{X}_t))_{kk}^2 \right) dt$$
$$+ \nabla_x \phi(\widetilde{X}_t, \widetilde{Y}_t) \cdot \left(\sigma_1 D(\widetilde{X}_t - u_t) dB_t^1 + \sigma_2 D(\widetilde{X}_t - \widetilde{Y}_t) dB_t^2 + \sigma_3 D(\nabla\mathcal{E}(\widetilde{X}_t)) dB_t^3 \right)$$

After taking the expectation, applying Fubini's theorem and observing that the stochastic integrals of the form $\mathbb{E} \int_0^t \nabla_x \phi(\widetilde{X}_t, \widetilde{Y}_t) \cdot D(\,\cdot\,) dB_t$ vanish as a consequence of [50, Theorem 3.2.1(iii)] due to the established regularity $\widetilde{\rho} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ and $\phi \in \mathcal{C}_*^2(\mathbb{R}^d \times \mathbb{R}^d)$, we obtain

$$\frac{d}{dt} \mathbb{E}\phi(\widetilde{X}_t, \widetilde{Y}_t) = -\mathbb{E}\nabla_x \phi(\widetilde{X}_t, \widetilde{Y}_t) \cdot \left(\lambda_1(\widetilde{X}_t - u_t) + \lambda_2(\widetilde{X}_t - \widetilde{Y}_t) + \lambda_3 \nabla\mathcal{E}(\widetilde{X}_t) \right)$$
$$+ \mathbb{E}\nabla_y \phi(\widetilde{X}_t, \widetilde{Y}_t) \cdot \left(\kappa(\widetilde{X}_t - \widetilde{Y}_t) S^{\beta,\theta}(\widetilde{X}_t, \widetilde{Y}_t) \right)$$
$$+ \frac{1}{2} \sum_{k=1}^d \partial^2_{x_k x_k} \phi(\widetilde{X}_t, \widetilde{Y}_t) \left(\sigma_1^2 D(\widetilde{X}_t - u_t)_{kk}^2 + \sigma_2^2 D(\widetilde{X}_t - \widetilde{Y}_t)_{kk}^2 + \sigma_3^2 D(\nabla\mathcal{E}(\widetilde{X}_t))_{kk}^2 \right)$$

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

$$\frac{d}{dt} \iint \phi(x, y) \, d\widetilde{\rho}_t(x, y) = -\iint \kappa S^{\beta,\theta}(x, y) \langle y - x, \nabla_y \phi(x, y) \rangle \, d\widetilde{\rho}_t(x, y)$$
$$- \iint \lambda_1 \langle x - u_t, \nabla_x \phi(x, y) \rangle + \lambda_2 \langle x - y, \nabla_x \phi(x, y) \rangle + \lambda_3 \langle \nabla\mathcal{E}(x), \nabla_x \phi(x, y) \rangle \, d\widetilde{\rho}_t(x, y) \quad (2.7)$$
$$+ \frac{1}{2} \iint \sum_{k=1}^d \left(\sigma_1^2 D(x - u_t)_{kk}^2 + \sigma_2^2 D(x - y)_{kk}^2 + \sigma_3^2 D(\nabla\mathcal{E}(x))_{kk}^2 \right) \partial^2_{x_k x_k} \phi(x, y) \, d\widetilde{\rho}_t(x, y)$$

The remainder is identical to the cited reference and is summarised below for completeness.

Step 3: Setting $\mathcal{T}u := y_\alpha(\widetilde{\rho}_Y) \in \mathcal{C}([0, T], \mathbb{R}^d)$ provides the self-mapping property of the map

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

which is compact as a consequence of a stability estimate for the consensus point [12, Lemma 3.2]. More precisely, as shown in the cited result, it holds $\|y_\alpha(\widetilde{\rho}_{Y,t}) - y_\alpha(\widetilde{\rho}_{Y,s})\|_2 \lesssim W_2(\widetilde{\rho}_{Y,t}, \widetilde{\rho}_{Y,s})$ for $\widetilde{\rho}_{Y,t}, \widetilde{\rho}_{Y,s} \in \mathcal{P}_4(\mathbb{R}^d)$. Together with the Hölder-1/2 continuity of the Wasserstein-2 distance $W_2(\widetilde{\rho}_{Y,t}, \widetilde{\rho}_{Y,s})$, this ensures the claimed compactness of $\mathcal{T}$.

Step 4: Then, for $u = \vartheta \mathcal{T}u$ with $\vartheta \in [0, 1]$, there exists $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ satisfying (2.7) with marginal ρ_Y such that $u = \vartheta y_\alpha(\rho_Y)$. For such u, a uniform bound can be obtained either thanks to the boundedness or the growth condition of $\mathcal{E}$ required in the statement. An application of the Leray-Schauder fixed point theorem concludes the proof by providing a solution to (1.5). □

2.2. Main result

We now present the main theoretical result about global mean-field law convergence of CBO with memory effects and gradient information for objectives that satisfy the following conditions.

Definition 2.4 (*Assumptions*). *Throughout, we are interested in functions* $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$, *for which*

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

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

$$\|x - x^*\|_\infty \leq \frac{1}{\eta}(\mathcal{E}(x) - \underline{\mathcal{E}})^\nu \quad \text{for all } x \in B_{R_0}^\infty(x^*), \quad (2.8)$$

$$\mathcal{E}_\infty < \mathcal{E}(x) - \underline{\mathcal{E}} \quad \text{for all } x \in \left(B_{R_0}^\infty(x^*) \right)^c. \quad (2.9)$$

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

A3 there exist $C_{\nabla\mathcal{E}} > 0$ such that

$$\|\nabla\mathcal{E}(x)\|_2 \leq C_{\nabla\mathcal{E}} \|x - x^*\|_2 \quad \text{for all } x \in \mathbb{R}^d. \tag{2.10}$$

In the case, where no gradient drift is present, i.e., $\lambda_3 = 0$ in equations (1.1), (1.5) and (1.6), the objective function $\mathcal{E}$ is only required to be continuous and satisfy Assumptions A1 and A2. While the former merely imposes that the infimum is attained at x^*, the latter can be regarded as a tractability condition of the energy landscape of $\mathcal{E}$ [26, 28]. More precisely, the inverse continuity condition (2.8) ensures that $\mathcal{E}$ is locally coercive in some neighbourhood of the global minimiser x^*. Condition (2.9), on the other hand, guarantees that in the farfield $\mathcal{E}$ is bounded away from the minimal value by at least $\mathcal{E}_\infty$. This in particular excludes objectives for which $\mathcal{E}(x) \approx \underline{\mathcal{E}}$ far away from x^*. Note that A2 actually already implies the uniqueness of x^* requested in A1. In case of an additional gradient drift term in the dynamics, i.e., $\lambda_3 \neq 0$, the objective naturally needs to be continuously differentiable. Furthermore, in Assumption A3 we impose that the gradient $\nabla\mathcal{E}$ grows at most linearly. This is a significantly weaker assumption compared to typical smoothness assumptions about $\mathcal{E}$ in the optimisation literature (in particular in the analysis of stochastic gradient descent), where Lipschitz-continuity of the gradient of $\mathcal{E}$ is required [49].

We are now ready to state the main theoretical result. Its proof is deferred to Section 3. For the reader's convenience let us recall that

$$W_2^2\left(\rho_t, \delta_{(x^*,x^*)}\right) = \iint \left(\|x - x^*\|_2^2 + \|y - x^*\|_2^2\right) d\rho_t(x, y),$$

which motivates to investigate the behaviour of the Lyapunov functional $\mathcal{V}(\rho_t)$ as introduced in (2.11) below.

Theorem 2.5. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 and A2. Furthermore, in the case of an active gradient drift in the CBO dynamcis (1.5), i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ obey in addition A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d)$ be such that $(x^*, x^*) \in \text{supp}(\rho_0)$. Let us define the functional*

$$\mathcal{V}(\rho_t) := \frac{1}{2} \iint \left(\|x - x^*\|_2^2 + \|y - x\|_2^2\right) d\rho_t(x, y), \tag{2.11}$$

and the rates

$$\chi_1 := \min\left\{\lambda_1 - \lambda_2 - 3\lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_1^2 - 2\sigma_3^2 C_{\nabla\mathcal{E}}^2, 2\kappa\theta + \lambda_2 - \lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_2^2\right\}, \tag{2.12a}$$

$$\chi_2 := \max\left\{3\lambda_1 + \lambda_2 + 3\lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_1^2 + 2\sigma_3^2 C_{\nabla\mathcal{E}}^2, 2\kappa\theta + 3\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - 2\sigma_2^2\right\}, \tag{2.12b}$$

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

$$T^* := \frac{1}{(1 - \vartheta)\chi_1} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right). \tag{2.13}$$

Then, there exists $\alpha_0 > 0$, depending (among problem-dependent quantities) also on ε and ϑ, such that for all $\alpha > \alpha_0$, if $\rho \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (1.6) on the time interval $[0, T^*]$ with initial condition ρ_0, we have

$$\mathcal{V}(\rho_T) = \varepsilon \quad \text{with} \quad T \in \left[\frac{(1 - \vartheta)\chi_1}{(1 + \vartheta/2)\chi_2} T^*, T^*\right]. \tag{2.14}$$

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

$$W_2^2(\rho_t, \delta_{(x^*,x^*)}) \leq 6\mathcal{V}(\rho_t) \leq 6\mathcal{V}(\rho_0) \exp\left(-(1 - \vartheta)\chi_1 t\right). \tag{2.15}$$

Theorem 2.5 proves the exponentially fast convergence of the law ρ of the dynamics (1.5) to the global minimiser x^* of $\mathcal{E}$ under a minimal assumption about the initial distribution ρ_0. The result in particular allows to devise a strategy for the parameter choices of the method. Namely, fixing the parameters $\lambda_2, \lambda_3, \sigma_1, \sigma_2, \sigma_3$ and θ, choosing λ_1 and consecutively κ such that

$$\lambda_1 > \lambda_2 + 3\lambda_3 C_{\nabla\mathcal{E}} + 2\sigma_1^2 + 2\sigma_3^2 C_{\nabla\mathcal{E}}^2 \quad \text{and} \quad \kappa > \frac{1}{2\theta}\left(-\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} + 2\sigma_2^2\right)$$

ensures that the convergence rate χ_1 is strictly positive. Since $\chi_2 \geq \chi_1$, $\chi_2 > 0$ as well. Given a desired accuracy ε, by consulting the proof in Section 3.4, we can further derive an estimate on the lower bound of α, namely

$$\alpha_0 \sim d + \log 16d + \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) - \log c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right) - \log \rho_0\left(B_r^\infty(x^*) \times B_r^\infty(x^*)\right)$$

for some suitably small $r \in (0, R_0)$, which, like the hidden constant, may depend on ε. The choice of the first set of parameters, in particular what concerns the drift towards the historical best and in the direction of the negative gradient, requires some manual hyperparameter tuning and depends on the problem at hand. We will see this also in Section 4, where we conduct numerical experiments in different application areas.

Eventually, with (2.13) one can determine the maximal time horizon T^*, until which the Lyapunov functional $\mathcal{V}(\rho_t)$ is guaranteed to have reached the prescribed ε. The exact time point T, where $\mathcal{V}(\rho_T) = \varepsilon$, is characterised more concretely in equation (2.14). Due to the presence of memory effects and gradient information, which might counteract the consensus drift of CBO, it seems challenging to specify T more closely. However, in the case of standard CBO, T turns out to be equal to T^* up to a factor depending merely on ϑ, see, e.g., [28].

In fact, this result can be retrieved as a special case of the subsequent Corollary 2.6, where we state an analogous convergence result for the CBO dynamics with gradient information but without memory effect. Its respective proof follows the lines of the one of the richer dynamics in Section 3, cf. also [28, Theorem 12] and [29, Theorem 2], and it is left as an exercise to the reader interested in the technical details of the proof technique. More precisely, for the instantaneous CBO model with gradient drift,

$$\begin{aligned}
d\widetilde{X}_t^i = {}&-\lambda_1\big(\widetilde{X}_t^i - y_\alpha(\widehat{\widetilde{\rho}}_t^N)\big)\,dt - \lambda_3 \nabla\mathcal{E}(\widetilde{X}_t^i)\,dt \\
&+ \sigma_1 D\big(\widetilde{X}_t^i - y_\alpha(\widehat{\widetilde{\rho}}_t^N)\big)\,d\widetilde{B}_t^{1,i} + \sigma_3 D\big(\nabla\mathcal{E}(\widetilde{X}_t^i)\big)\,d\widetilde{B}_t^{3,i},
\end{aligned} \tag{2.16}$$

where $\widehat{\widetilde{\rho}}_t^N := \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}_t^i}$ and to which the associated mean-field Fokker-Planck equation reads

$$\begin{aligned}
\partial_t \widetilde{\rho}_t = {}&\operatorname{div}\big(\big(\lambda_1\,(x - y_\alpha(\widetilde{\rho}_t)) + \lambda_3 \nabla\mathcal{E}(x)\big)\widetilde{\rho}_t\big) \\
&+ \frac{1}{2}\sum_{k=1}^d \partial_{x_k x_k}^2\left(\big(\sigma_1^2 D(x - y_\alpha(\widetilde{\rho}_t))_{kk}^2 + \sigma_3^2 D(\nabla\mathcal{E}(x))_{kk}^2\big)\widetilde{\rho}_t\right),
\end{aligned} \tag{2.17}$$

we have the following convergence result.

Corollary 2.6. *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 and A2. Furthermore, in the case of an active gradient drift, i.e., if $\lambda_3 \neq 0$, let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ obey in addition A3. Moreover, let $\widetilde{\rho}_0 \in \mathcal{P}_4(\mathbb{R}^d)$ be such that $x^* \in \operatorname{supp}(\widetilde{\rho}_0)$. Let us define the functional*

$$\widetilde{\mathcal{V}}(\widetilde{\rho}_t) := \frac{1}{2}\int \|x - x^*\|_2^2\,d\widetilde{\rho}_t(x), \tag{2.18}$$

and the rates

$$\widetilde{\chi}_1 := 2\lambda_1 - 2\lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla\mathcal{E}}^2, \tag{2.19a}$$

$$\widetilde{\chi}_2 := 2\lambda_1 + 2\lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2, \tag{2.19b}$$

which we assume to be strictly positive through a sufficient choice of the parameters of the CBO dynamics. Furthermore, provided that $\widetilde{\mathcal{V}}(\widetilde{\rho}_0) > 0$, fix any $\varepsilon \in (0, \widetilde{\mathcal{V}}(\widetilde{\rho}_0))$, $\vartheta \in (0, 1)$ and define the time horizon

$$\widetilde{T}^* := \frac{1}{(1-\vartheta)\widetilde{\chi}_1} \log\left(\frac{\widetilde{\mathcal{V}}(\widetilde{\rho}_0)}{\varepsilon}\right). \tag{2.20}$$

Then, there exists $\widetilde{\alpha}_0 > 0$, depending (among problem-dependent quantities) also on ε and ϑ, such that for all $\alpha > \widetilde{\alpha}_0$, if $\widetilde{\rho} \in \mathcal{C}([0, T^*], \mathcal{P}_4(\mathbb{R}^d))$ is a weak solution to the Fokker-Planck equation (2.17) on the time interval $[0, \widetilde{T}^*]$ with initial condition $\widetilde{\rho}_0$, we have

$$\widetilde{\mathcal{V}}(\widetilde{\rho}_{\widetilde{T}}) = \varepsilon \quad with \quad \widetilde{T} \in \left[\frac{(1-\vartheta)\widetilde{\chi}_1}{(1+\vartheta/2)\widetilde{\chi}_2}\widetilde{T}^*, \widetilde{T}^*\right]. \tag{2.21}$$

Furthermore, on the time interval $[0, \widetilde{T}]$, $\widetilde{\mathcal{V}}(\widetilde{\rho}_t)$ decays at least exponentially fast, i.e., for all $t \in [0, \widetilde{T}]$ it holds

$$W_2^2(\widetilde{\rho}_t, \delta_{x^*}) = 2\widetilde{\mathcal{V}}(\widetilde{\rho}_t) \leq 2\widetilde{\mathcal{V}}(\widetilde{\rho}_0) \exp\left(-(1-\vartheta)\widetilde{\chi}_1 t\right). \tag{2.22}$$

3. Proof details for Section 2.2

In what follows, we provide the proof details for the global mean-field law convergence result of CBO with memory effects and gradient information, Theorem 2.5. The entire section can be read as a proof sketch with Corollaries 3.3 and 3.5, Propositions 3.6 and 3.8 containing the key individual statements. How to combine these results rigorously to complete the proof of Theorem 2.5 is then covered in detail in Section 3.4.

Remark 3.1. *Without loss of generality, we assume $\underline{\mathcal{E}} = 0$ throughout this section.*

3.1. Evolution of the mean-field limit

Recall that our overall goal is to establish the convergence of the dynamics (1.6) to a Dirac delta at the global minimiser x^* with respect to the Wasserstein-2 distance, i.e.,

$$W_2(\rho_t, \delta_{(x^*, x^*)}) \to 0 \quad as \quad t \to \infty.$$

To this end, we analyse the decay behaviour of the functional $\mathcal{V}(\rho_t)$ as defined in (2.11), i.e., $\mathcal{V}(\rho_t) = \frac{1}{2}\iint\left(\|x - x^*\|_2^2 + \|y - x\|_2^2\right)d\rho_t(x, y)$. More precisely, we will show its exponential decay with a rate controllable through the parameters of the CBO method.

Let us start below with deriving the evolution inequalities for the functionals

$$\mathcal{X}(\rho_t) = \frac{1}{2}\iint \|x - x^*\|_2^2\, d\rho_t(x, y) \quad and \quad \mathcal{Y}(\rho_t) = \frac{1}{2}\iint \|y - x\|_2^2\, d\rho_t(x, y).$$

Lemma 3.2. *Let $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$, and fix $\alpha, \lambda_1, \sigma_1 > 0$ and $\lambda_2, \sigma_2, \lambda_3, \sigma_3, \beta, \kappa, \theta \geq 0$. Moreover, let $T > 0$ and let $\rho \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^d \times \mathbb{R}^d))$ be a weak solution to the Fokker-Planck equation (1.6). Then, the functionals $\mathcal{X}(\rho_t)$ and $\mathcal{Y}(\rho_t)$ satisfy*

$$\frac{d}{dt}\begin{pmatrix}\mathcal{X}(\rho_t)\\\mathcal{Y}(\rho_t)\end{pmatrix} \leq -\begin{pmatrix}2\lambda_1 - \lambda_2 - 2\lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla\mathcal{E}}^2 & -\lambda_2 - \sigma_2^2\\ -\lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla\mathcal{E}}^2 & 2\kappa\theta + 2\lambda_2 - \lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - \sigma_2^2\end{pmatrix}\begin{pmatrix}\mathcal{X}(\rho_t)\\\mathcal{Y}(\rho_t)\end{pmatrix}$$
$$+ \sqrt{2}\begin{pmatrix}(\lambda_1 + \sigma_1^2)\sqrt{\mathcal{X}(\rho_t)}\\\lambda_1\sqrt{\mathcal{Y}(\rho_t)} + \sigma_1^2\sqrt{\mathcal{X}(\rho_t)}\end{pmatrix}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2 + \frac{\sigma_1^2}{2}\begin{pmatrix}1\\1\end{pmatrix}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2^2,$$

where the inequality has to be understood component-wise.

Proof. We note that the functions $\phi_{\mathcal{X}}(x, y) = 1/2 \, \|x - x^*\|_2^2$ and $\phi_{\mathcal{Y}}(x, y) = 1/2 \, \|y - x\|_2^2$ are in $\mathcal{C}_*^2(\mathbb{R}^d \times \mathbb{R}^d)$ and recall that ρ satisfies the weak solution identity (2.1) for such test functions. Hence, by applying (2.1) with $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ as above, we obtain for the evolution of $\mathcal{X}(\rho_t)$

$$
\begin{aligned}
\frac{d}{dt} \mathcal{X}(\rho_t) = & -\iint \lambda_1 \langle x - y_\alpha(\rho_{Y,t}), x - x^* \rangle + \lambda_2 \langle x - y, x - x^* \rangle + \lambda_3 \langle \nabla \mathcal{E}(x), x - x^* \rangle \, d\rho_t(x, y) \\
& + \frac{1}{2} \iint \sigma_1^2 \, \|x - y_\alpha(\rho_{Y,t})\|_2^2 + \sigma_2^2 \, \|x - y\|_2^2 + \sigma_3^2 \, \|\nabla \mathcal{E}(x)\|_2^2 \, d\rho_t(x, y)
\end{aligned}
\tag{3.1}
$$

and for the evolution of $\mathcal{Y}(\rho_t)$

$$
\begin{aligned}
\frac{d}{dt} \mathcal{Y}(\rho_t) = & -\iint \kappa S^{\beta,\theta}(x, y) \, \|x - y\|_2^2 \, d\rho_t(x, y) \\
& -\iint \lambda_1 \langle x - y_\alpha(\rho_{Y,t}), x - y \rangle + \lambda_2 \, \|x - y\|_2^2 + \lambda_3 \langle \nabla \mathcal{E}(x), x - y \rangle \, d\rho_t(x, y) \\
& + \frac{1}{2} \iint \sigma_1^2 \, \|x - y_\alpha(\rho_{Y,t})\|_2^2 + \sigma_2^2 \, \|x - y\|_2^2 + \sigma_3^2 \, \|\nabla \mathcal{E}(x)\|_2^2 \, d\rho_t(x, y).
\end{aligned}
\tag{3.2}
$$

Here we used $\nabla_x \phi_{\mathcal{X}}(x, y) = x - x^*$, $\nabla_y \phi_{\mathcal{X}}(x, y) = 0$, $\partial_{x_k x_k}^2 \phi_{\mathcal{X}}(x, y) = 1$, $\nabla_x \phi_{\mathcal{Y}}(x, y) = x - y$, $\nabla_y \phi_{\mathcal{Y}}(x, y) = y - x$ and $\partial_{x_k x_k}^2 \phi_{\mathcal{Y}}(x, y) = 1$. Let us now collect auxiliary estimates in (3.3a)–(3.3g), which turn out to be useful in establishing upper bounds for (3.1) and (3.2). Using standard tools such as Cauchy-Schwarz and Young's inequality we have

$$
-\langle x - y, x - x^* \rangle \le \|x - y\|_2 \|x - x^*\|_2 \le \frac{1}{2} \left(\|x - y\|_2^2 + \|x - x^*\|_2^2 \right),
\tag{3.3a}
$$

$$
\begin{aligned}
-\langle x - y_\alpha(\rho_{Y,t}), x - x^* \rangle & = -\|x - x^*\|_2^2 - \langle x^* - y_\alpha(\rho_{Y,t}), x - x^* \rangle \\
& \le -\|x - x^*\|_2^2 + \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - x^*\|_2,
\end{aligned}
\tag{3.3b}
$$

$$
\begin{aligned}
-\langle x - y_\alpha(\rho_{Y,t}), x - y \rangle & = -\langle x - x^*, x - y \rangle - \langle x^* - y_\alpha(\rho_{Y,t}), x - y \rangle \\
& \le \frac{1}{2} \left(\|x - y\|_2^2 + \|x - x^*\|_2^2 \right) + \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - y\|_2,
\end{aligned}
\tag{3.3c}
$$

$$
\begin{aligned}
\|x - y_\alpha(\rho_{Y,t})\|_2^2 & = \|x - x^*\|_2^2 - 2 \langle y_\alpha(\rho_{Y,t}) - x^*, x - x^* \rangle + \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2 \\
& \le \|x - x^*\|_2^2 + 2 \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \|x - x^*\|_2 + \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2,
\end{aligned}
\tag{3.3d}
$$

where in (3.3b)–(3.3d) we expanded the left-hand side of the scalar product and the norm by subtracting and adding x^*. Furthermore, by means of A3 we obtain

$$
-\langle \nabla \mathcal{E}(x), x - x^* \rangle \le \|\nabla \mathcal{E}(x)\|_2 \|x - x^*\|_2 \le C_{\nabla \mathcal{E}} \|x - x^*\|_2^2,
\tag{3.3e}
$$

$$
\begin{aligned}
-\langle \nabla \mathcal{E}(x), x - y \rangle & \le \|\nabla \mathcal{E}(x)\|_2 \|x - y\|_2 \le C_{\nabla \mathcal{E}} \|x - x^*\|_2 \|x - y\|_2 \\
& \le \frac{C_{\nabla \mathcal{E}}}{2} \left(\|x - y\|_2^2 + \|x - x^*\|_2^2 \right),
\end{aligned}
\tag{3.3f}
$$

$$
\|\nabla \mathcal{E}(x)\|_2^2 \le C_{\nabla \mathcal{E}}^2 \|x - x^*\|_2^2.
\tag{3.3g}
$$

Integrating the bounds (3.3a), (3.3b), (3.3d), (3.3e) and (3.3g) into equation (3.1) results in the upper bound

$$
\begin{aligned}
\frac{d}{dt} \mathcal{X}(\rho_t) \le & -\left(2\lambda_1 - \lambda_2 - 2\lambda_3 C_{\nabla \mathcal{E}} - \sigma_1^2 - \sigma_3^2 C_{\nabla \mathcal{E}}^2 \right) \mathcal{X}(\rho_t) + \left(\lambda_2 + \sigma_2^2 \right) \mathcal{Y}(\rho_t) \\
& + \sqrt{2} \left(\lambda_1 + \sigma_1^2 \right) \sqrt{\mathcal{X}(\rho_t)} \, \|y_\alpha(\rho_{Y,t}) - x^*\|_2 + \frac{\sigma_1^2}{2} \|y_\alpha(\rho_{Y,t}) - x^*\|_2^2,
\end{aligned}
$$

where we furthermore used that by Jensen's inequality

$$
\iint \|x - x^*\|_2 \, d\rho_t(x, y) \le \sqrt{\iint \|x - x^*\|_2^2 \, d\rho_t(x, y)} = \sqrt{2 \mathcal{X}(\rho_t)}.
\tag{3.4}
$$

For equation (3.2), we first note that, by definition, $S^{\beta,\theta} \geq \theta$ uniformly. This combined with the bounds (3.3c), (3.3d), (3.3f) and (3.3g) allows to derive

$$\frac{d}{dt}\mathcal{Y}(\rho_t) \leq -\left(2\kappa\theta + 2\lambda_2 - \lambda_1 - \lambda_3 C_{\nabla\mathcal{E}} - \sigma_2^2\right)\mathcal{Y}(\rho_t) + \left(\lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} + \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2\right)\mathcal{X}(\rho_t)$$
$$+ \sqrt{2}\left(\lambda_1\sqrt{\mathcal{Y}(\rho_t)} + \sigma_1^2\sqrt{\mathcal{X}(\rho_t)}\right)\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2 + \frac{\sigma_1^2}{2}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2^2,$$

where we used (3.4) together with an analogous bound for $\iint \|x - y\|_2 \, d\rho_t(x, y)$.   □

Recalling that $\mathcal{V}(\rho_t) = \mathcal{X}(\rho_t) + \mathcal{Y}(\rho_t)$ immediately allows to obtain an evolution inequality for $\mathcal{V}(\rho_t)$ of the following form.

Corollary 3.3. *Under the assumptions of Lemma 3.2, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt}\mathcal{V}(\rho_t) \leq -\chi_1 \mathcal{V}(\rho_t) + 2\sqrt{2}\left(\lambda_1 + \sigma_1^2\right)\sqrt{\mathcal{V}(\rho_t)}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2 + \sigma_1^2\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2^2, \tag{3.5}$$

with χ_1 as specified in (2.12a).

Analogously to the upper bounds on the time evolutions of the functionals $\mathcal{X}(\rho_t)$, $\mathcal{Y}(\rho_t)$ and $\mathcal{V}(\rho_t)$, we can derive bounds from below as follows.

Lemma 3.4. *Under the assumptions of Lemma 3.2, the functionals $\mathcal{X}(\rho_t)$ and $\mathcal{Y}(\rho_t)$ satisfy*

$$\frac{d}{dt}\begin{pmatrix}\mathcal{X}(\rho_t)\\\mathcal{Y}(\rho_t)\end{pmatrix} \geq -\begin{pmatrix}2\lambda_1 + \lambda_2 + 2\lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2 & \lambda_2 - \sigma_2^2\\ \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - \sigma_1^2 + \sigma_3^2 C_{\nabla\mathcal{E}}^2 & 2\kappa\theta + 2\lambda_2 + \lambda_1 + \lambda_3 C_{\nabla\mathcal{E}} - \sigma_2^2\end{pmatrix}\begin{pmatrix}\mathcal{X}(\rho_t)\\\mathcal{Y}(\rho_t)\end{pmatrix}$$
$$- \sqrt{2}\begin{pmatrix}\left(\lambda_1 + \sigma_1^2\right)\sqrt{\mathcal{X}(\rho_t)}\\ \lambda_1\sqrt{\mathcal{Y}(\rho_t)} + \sigma_1^2\sqrt{\mathcal{X}(\rho_t)}\end{pmatrix}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2,$$

where the inequality has to be understood component-wise.

Proof. By following the lines of the proof of Lemma 3.2 and noticing that in analogy to the estimates (3.3), it hold

$$-\langle x - y, x - x^*\rangle \geq -\|x - y\|_2\|x - x^*\|_2 \geq -\frac{1}{2}\left(\|x - y\|_2^2 + \|x - x^*\|_2^2\right), \tag{3.6a}$$

$$-\langle x - y_\alpha(\rho_{Y,t}), x - x^*\rangle = -\|x - x^*\|_2^2 - \langle x^* - y_\alpha(\rho_{Y,t}), x - x^*\rangle$$
$$\geq -\|x - x^*\|_2^2 - \|y_\alpha(\rho_{Y,t}) - x^*\|_2\|x - x^*\|_2, \tag{3.6b}$$

$$-\langle x - y_\alpha(\rho_{Y,t}), x - y\rangle = -\langle x - x^*, x - y\rangle - \langle x^* - y_\alpha(\rho_{Y,t}), x - y\rangle$$
$$\geq -\frac{1}{2}\left(\|x - y\|_2^2 + \|x - x^*\|_2^2\right) - \|y_\alpha(\rho_{Y,t}) - x^*\|_2\|x - y\|_2, \tag{3.6c}$$

$$\left\|x - y_\alpha(\rho_{Y,t})\right\|_2^2 = \|x - x^*\|_2^2 - 2\langle y_\alpha(\rho_{Y,t}) - x^*, x - x^*\rangle + \left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2^2$$
$$\geq \|x - x^*\|_2^2 - 2\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2\|x - x^*\|_2, \tag{3.6d}$$

as well as

$$-\langle\nabla\mathcal{E}(x), x - x^*\rangle \geq -\|\nabla\mathcal{E}(x)\|_2\|x - x^*\|_2 \geq -C_{\nabla\mathcal{E}}\|x - x^*\|_2^2, \tag{3.6e}$$

$$-\langle\nabla\mathcal{E}(x), x - y\rangle \geq -\|\nabla\mathcal{E}(x)\|_2\|x - y\|_2 \geq -C_{\nabla\mathcal{E}}\|x - x^*\|_2\|x - y\|_2$$
$$\geq -\frac{C_{\nabla\mathcal{E}}}{2}\left(\|x - y\|_2^2 + \|x - x^*\|_2^2\right), \tag{3.6f}$$

$$\|\nabla\mathcal{E}(x)\|_2^2 \geq -C_{\nabla\mathcal{E}}^2\|x - x^*\|_2^2. \tag{3.6g}$$

we obtain the statement by integrating the bounds into equations (3.1) and (3.2).   □

Corollary 3.5. *Under the assumptions of Lemma 3.2, the functional $\mathcal{V}(\rho_t)$ satisfies*

$$\frac{d}{dt}\mathcal{V}(\rho_t) \geq -\chi_2 \mathcal{V}(\rho_t) - 2\sqrt{2}\left(\lambda_1 + \sigma_1^2\right)\sqrt{\mathcal{V}(\rho_t)}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2, \tag{3.7}$$

with χ_2 as specified in (2.12b).

In order to be able to apply Grönwall's inequality to (3.5) and (3.7) with the aim of obtaining estimates of the form $\mathcal{V}(\rho_t) \leq \mathcal{V}(\rho_0)e^{-(1-\vartheta)\chi_1 t}$ and $\mathcal{V}(\rho_t) \geq \mathcal{V}(\rho_0)e^{-(1-\vartheta/2)\chi_2 t}$ for some $\chi_1, \chi_2 > 0$ and a suitable $\vartheta \in (0, 1)$, it remains to control the quantity $\left\| y_\alpha(\rho_{Y,t}) - x^* \right\|_2$ through the choice of the parameter α. This is the content of the next section.

3.2. Quantitative Laplace principle

The well-known Laplace principle [21, 48, 52] asserts that for any absolutely continuous probability distribution $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with $x^* \in \operatorname{supp}(\varrho)$ it holds

$$\lim_{\alpha \to \infty} \left(-\frac{1}{\alpha} \log \|\omega_\alpha\|_{L_1(\varrho)} \right) = \mathcal{E}(x^*) = \underline{\mathcal{E}}, \tag{3.8}$$

which allows to infer that the α-weighted measure $\omega_\alpha / \|\omega_\alpha\|_{L_1(\varrho)} \varrho$ is concentrated in a small region around the minimiser x^*, provided that $\mathcal{E}$ attains its minimum at a single point, which is however guaranteed by the inverse continuity property A2.

The asymptotic nature of the result (3.8), however, does not permit to obtain the required quantitative estimates, which is the reason why the authors of [28] proposed a quantitative nonasymptotic variant of the Laplace principle. In the following proposition, cf. [29, Proposition 1], we state this result for the setting of anisotropic noise considered throughout the paper.

Proposition 3.6 *([29, Proposition 1]). Let $\underline{\mathcal{E}} = 0$, $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and fix $\alpha > 0$. For any $r > 0$ we define $\mathcal{E}_r := \sup_{y \in B_r^\infty(x^*)} \mathcal{E}(y)$. Then, under the inverse continuity property A2, for any $r \in (0, R_0]$ and $q > 0$ such that $q + \mathcal{E}_r \leq \mathcal{E}_\infty$, we have*

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

Proof. The proof is a mere reformulation of the one of [29, Proposition 1], which is presented in what follows for the sake of completeness.

For any $a > 0$, Markov's inequality gives $\|\omega_\alpha\|_{L_1(\varrho)} \geq a\varrho(\{y : \exp(-\alpha \mathcal{E}(y)) \geq a\})$. By choosing $a = \exp(-\alpha \mathcal{E}_r)$ and noting that

$$\varrho\left(\{y \in \mathbb{R}^d : \exp(-\alpha \mathcal{E}(y)) \geq \exp(-\alpha \mathcal{E}_r)\}\right) = \varrho\left(\{y \in \mathbb{R}^d : \mathcal{E}(y) \leq \mathcal{E}_r\}\right) \geq \varrho(B_r^\infty(x^*)),$$

we get $\|\omega_\alpha\|_{L_1(\varrho)} \geq \exp(-\alpha \mathcal{E}_r)\varrho(B_r^\infty(x^*))$. Now let $\widetilde{r} \geq r > 0$. With the definition of the consensus point $y_\alpha(\varrho) = \int y\omega_\alpha(y)/\|\omega_\alpha\|_{L_1(\varrho)} \, d\varrho(y)$ and Jensen's inequality, we can decompose

$$\|y_\alpha(\varrho) - x^*\|_\infty \leq \int_{B_{\widetilde{r}}^\infty(x^*)} \|y - x^*\|_\infty \frac{\omega_\alpha(y)}{\|\omega_\alpha\|_{L_1(\varrho)}} \, d\varrho(y)$$
$$+ \int_{\left(B_{\widetilde{r}}^\infty(x^*)\right)^c} \|y - x^*\|_\infty \frac{\omega_\alpha(y)}{\|\omega_\alpha\|_{L_1(\varrho)}} \, d\varrho(y).$$

After noticing that the first term is bounded by $\widetilde{r}$ since $\|y - x^*\|_\infty \leq \widetilde{r}$ for all $y \in B_{\widetilde{r}}^\infty(x^*)$, we can continue the former with

$$\|y_\alpha(\varrho) - x^*\|_\infty \leq \widetilde{r} + \frac{1}{\exp(-\alpha \mathcal{E}_r)\varrho(B_r^\infty(x^*))} \int_{(B_{\widetilde{r}}^\infty(x^*))^c} \|y - x^*\|_\infty \omega_\alpha(y) \, d\varrho(y)$$
$$\leq \widetilde{r} + \frac{\exp\left(-\alpha \inf_{y \in (B_{\widetilde{r}}^\infty(x^*))^c} \mathcal{E}(y)\right)}{\exp(-\alpha \mathcal{E}_r)\varrho(B_r^\infty(x^*))} \int_{(B_{\widetilde{r}}^\infty(x^*))^c} \|y - x^*\|_\infty \, d\varrho(y)$$
$$= \widetilde{r} + \frac{\exp\left(-\alpha \left(\inf_{y \in (B_{\widetilde{r}}^\infty(x^*))^c} \mathcal{E}(y) - \mathcal{E}_r\right)\right)}{\varrho(B_r^\infty(x^*))} \int \|y - x^*\|_\infty \, d\varrho(y), \tag{3.9}$$

where for the second term we used $\|\omega_\alpha\|_{L_1(\varrho)} \geq \exp(-\alpha\mathcal{E}_r)\varrho(B_r^\infty(x^*))$ from above. Let us now choose $\widetilde{r} = (q + \mathcal{E}_r)^\nu/\eta$, which satisfies $\widetilde{r} \geq r$, since A2 with $\underline{\mathcal{E}} = 0$ and $r \leq R_0$ implies

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

Furthermore, due to the assumption $q + \mathcal{E}_r \leq \mathcal{E}_\infty$ in the statement we have $\widetilde{r} \leq \mathcal{E}_\infty^\nu/\eta$, which together with the two cases of A2 with $\underline{\mathcal{E}} = 0$ allows to bound the infimum in (3.9) as follows

$$\inf_{y \in (B_r^\infty(x^*))^c} \mathcal{E}(y) - \mathcal{E}_r \geq \min\left\{\mathcal{E}_\infty, (\eta\widetilde{r})^{\frac{1}{\nu}}\right\} - \mathcal{E}_r = (\eta\widetilde{r})^{\frac{1}{\nu}} - \mathcal{E}_r = (q + \mathcal{E}_r) - \mathcal{E}_r = q.$$

Inserting this and the definition of $\widetilde{r}$ into (3.9), we get the result as a consequence of the norm equivalence $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \sqrt{d}\,\|\cdot\|_\infty$. $\qquad\square$

To eventually apply Proposition 3.6 in the setting of Corollary 3.3, i.e., to upper bound the distance of the consensus point $y_\alpha(\rho_{Y,t})$ to the global minimiser x^*, it remains to ensure that $\rho_{Y,t}(B_r^\infty(x^*))$ is bounded away from 0 for a finite time horizon. We ensure that this is indeed the case in what follows.

3.3. A lower bound for the probability mass $\rho_{Y,t}(B_r^\infty(x^*))$

In this section, for any small radius $r > 0$, we provide a lower bound on the probability mass of $\rho_{Y,t}(B_r^\infty(x^*))$ by defining a mollifier $\phi_r : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ so that

$$\rho_{Y,t}(B_r^\infty(x^*)) = \rho_t(\mathbb{R}^d \times B_r^\infty(x^*)) = \iint_{\mathbb{R}^d \times B_r^\infty(x^*)} 1\, d\rho_t(x, y) \geq \iint \phi_r(x, y)\, d\rho_t(x, y)$$

and studying the evolution of the right-hand side.

Lemma 3.7. *For $r > 0$ let $\Omega_r := \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \max\{\|x - x^*\|_\infty, \|x - y\|_\infty\} < r/2\}$ and define the mollifier $\phi_r : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by*

$$\phi_r(x, y) := \begin{cases} \prod_{k=1}^d \exp\left(1 - \frac{\left(\frac{r}{2}\right)^2}{\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2}\right) \exp\left(1 - \frac{\left(\frac{r}{2}\right)^2}{\left(\frac{r}{2}\right)^2 - (x - y)_k^2}\right), & \text{if } (x, y) \in \Omega_r, \\ 0, & \text{else.} \end{cases}$$

We have that $\mathrm{Im}(\phi_r) = [0, 1]$, $\mathrm{supp}(\phi_r) = \Omega_r \subset B_{r/2}^\infty(x^*) \times B_r^\infty(x^*) \subset \mathbb{R}^d \times B_r^\infty(x^*)$, $\phi_r \in \mathcal{C}_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ and

$$\partial_{x_k}\phi_r(x, y) = -\frac{r^2}{2}\left(\frac{(x - x^*)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2\right)^2} + \frac{(x - y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2}\right)\phi_r(x, y),$$

$$\partial_{y_k}\phi_r(x, y) = -\frac{r^2}{2}\frac{(y - x)_k}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2}\phi_r(x, y),$$

$$\partial_{x_k x_k}^2 \phi_r(x, y) = \frac{r^2}{2}\left(\left(\frac{2\left(2(x - x^*)_k^2 - \left(\frac{r}{2}\right)^2\right)(x - x^*)_k^2 - \left(\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2\right)^2}{\left(\left(\frac{r}{2}\right)^2 - (x - x^*)_k^2\right)^4}\right)\right.$$
$$\left. + \left(\frac{2\left(2(x - y)_k^2 - \left(\frac{r}{2}\right)^2\right)(x - y)_k^2 - \left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^2}{\left(\left(\frac{r}{2}\right)^2 - (x - y)_k^2\right)^4}\right)\right)\phi_r(x, y).$$

Proof. It is straightforward to check the properties of ϕ_r as it is a tensor product of classical well-studied mollifiers. $\qquad\square$

To keep the notation as concise as possible in what follows, let us introduce the decomposition

$$\partial_{x_k}\phi_r = \delta^*_{x_k}\phi_r + \delta^Y_{x_k}\phi_r \quad \text{and} \quad \partial^2_{x_k x_k}\phi_r = \delta^{2,*}_{x_k x_k}\phi_r + \delta^{2,Y}_{x_k x_k}\phi_r, \tag{3.10}$$

where

$$\delta^*_{x_k}\phi_r(x,y) = \frac{-\frac{r^2}{2}(x-x^*)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-x^*)^2_k\right)^2}\phi_r(x,y) \quad \text{and} \quad \delta^Y_{x_k}\phi_r(x,y) = \frac{-\frac{r^2}{2}(x-y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-y)^2_k\right)^2}\phi_r(x,y)$$

and analogously for $\delta^{2,*}_{x_k x_k}\phi_r$ and $\delta^{2,Y}_{x_k x_k}\phi_r$.

Proposition 3.8. *Let $T > 0$, $r > 0$, and fix parameters $\alpha, \lambda_1, \sigma_1 > 0$ as well as parameters $\lambda_2, \sigma_2, \lambda_3, \sigma_3, \beta, \kappa, \theta \geq 0$ such that $\sigma_2 > 0$ iff $\lambda_2 \neq 0$ and $\sigma_3 > 0$ iff $\lambda_3 \neq 0$. Moreover, assume the validity of Assumption A3 if $\lambda_3 \neq 0$. Let $\rho \in \mathcal{C}([0,T], \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d))$ weakly solve the Fokker-Planck equation (1.6) in the sense of Definition 2.1 with initial condition $\rho_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ and for $t \in [0,T]$. Then, for all $t \in [0,T]$ we have*

$$\rho_{Y,t}(B^\infty_r(x^*)) \geq \left(\iint \phi_r(x,y)\, d\rho_0(x,y)\right) \exp(-pt) \tag{3.11}$$

with

$$p := d \sum_{i=1}^3 \omega_i \left(\left(1 + \mathbb{1}_{i\in\{1,3\}}\right)\left(\frac{2\lambda_i C_\Upsilon \sqrt{c}}{(1-c)^2\frac{r}{2}} + \frac{\sigma_i^2 C_\Upsilon^2}{(1-c)^4\left(\frac{r}{2}\right)^2} + \frac{4\lambda_i^2}{\tilde{c}\sigma_i^2}\right) + \mathbb{1}_{i=2}\frac{\sigma_2^2 c}{(1-c)^4}\right), \tag{3.12}$$

where, for any $B < \infty$ with $\sup_{t\in[0,T]}\|y_\alpha(\rho_{Y,t}) - x^*\|_2 \leq B$, $C_\Upsilon = C_\Upsilon(r, B, d, C_{\nabla\mathcal{E}})$ is as defined in (3.20). Moreover, $\omega_i = \mathbb{1}_{\lambda_i > 0}$ for $i \in \{1,2,3\}$ and $c \in (1/2, 1)$ can be any constant that satisfies $(1-c)^2 \leq (2c-1)c$.

Remark 3.9. *In order to ensure a finite decay rate $p < \infty$ in Proposition 3.8, it is crucial to have non-vanishing diffusions $\sigma_1 > 0$, $\sigma_2 > 0$ if $\lambda_2 \neq 0$ and $\sigma_3 > 0$ if $\lambda_3 \neq 0$. As apparent from the formulation of the statement as well as the proof below, σ_2 or σ_3 may be 0 if the corresponding drift parameter, λ_2 or λ_3, respectively, vanishes.*

Proof of Proposition 3.8. By the definition of the marginal ρ_Y and the properties of the mollifier ϕ_r defined in Lemma 3.7, we have

$$\rho_{Y,t}(B^\infty_r(x^*)) = \rho_t(\mathbb{R}^d \times B^\infty_r(x^*)) \geq \rho_t(\Omega_r) \geq \iint \phi_r(x,y)\, d\rho_t(x,y).$$

Our strategy is to derive a lower bound for the right-hand side of this inequality. Using the weak solution property of ρ as in Definition 2.1 and the fact that $\phi_r \in \mathcal{C}_c^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, we obtain

$$\frac{d}{dt}\iint \phi_r(x,y)\, d\rho_t(x,y) = \sum_{k=1}^d \iint T^s_k(x,y)\, d\rho_t(x,y)$$

$$+ \sum_{k=1}^d \iint \left(T^c_{1k}(x,y) + T^c_{2k}(x,y) + T^\ell_{1k}(x,y) + T^\ell_{2k}(x,y) + T^g_{1k}(x,y) + T^g_{2k}(x,y)\right) d\rho_t(x,y), \tag{3.13}$$

where $T^s_k(x,y) := -\kappa S^{\beta,\theta}(x,y)(y-x)_k\, \partial_{y_k}\phi_r(x,y)$ and

$$T^c_{1k}(x,y) := -\lambda_1\left(x - y_\alpha(\rho_{Y,t})\right)_k \partial_{x_k}\phi_r(x,y), \quad T^c_{2k}(x,y) := \frac{\sigma_1^2}{2}\left(x - y_\alpha(\rho_{Y,t})\right)^2_k \partial^2_{x_k x_k}\phi_r(x,y),$$

$$T^\ell_{1k}(x,y) := -\lambda_2(x-y)_k\, \partial_{x_k}\phi_r(x,y), \quad T^\ell_{2k}(x,y) := \frac{\sigma_2^2}{2}(x-y)^2_k\, \partial^2_{x_k x_k}\phi_r(x,y),$$

$$T^g_{1k}(x,y) := -\lambda_3\partial_{x_k}\mathcal{E}(x)\partial_{x_k}\phi_r(x,y), \quad T^g_{2k}(x,y) := \frac{\sigma_3^2}{2}\left(\partial_{x_k}\mathcal{E}(x)\right)^2 \partial^2_{x_k x_k}\phi_r(x,y)$$

for $k \in \{1, \ldots, d\}$. Since the mollifier ϕ_r and its derivatives vanish outside of Ω_r, we restrict our attention to Ω_r and aim for showing for all $k \in \{1, \ldots, d\}$ that

- $T_k^s(x, y) \geq 0$,
- $T_{1k}^c(x, y) + T_{2k}^c(x, y) \geq -p^c \phi_r(x, y)$,
- $T_{1k}^\ell(x, y) + T_{2k}^\ell(x, y) \geq -p^\ell \phi_r(x, y)$,
- $T_{1k}^g(x, y) + T_{2k}^g(x, y) \geq -p^g \phi_r(x, y)$

pointwise for all $(x, y) \in \Omega_r$ with suitable constants $0 \leq p^\ell, p^c, p^g < \infty$.

Term T_k^s: Using the expression for $\partial_{y_k}\phi_r$ from Lemma 3.7 and the fact that $S^{\beta,\theta} \geq \theta/2 \geq 0$, it is easy to see that

$$T_k^s(x, y) = \frac{r^2\kappa}{2} S^{\beta,\theta}(x, y) \frac{(y-x)_k^2}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \phi_r(x, y) \geq 0. \tag{3.14}$$

Terms $T_{1k}^c + T_{2k}^c$, $T_{1k}^\ell + T_{2k}^\ell$ and $T_{1k}^g + T_{2k}^g$: We first note that the third inequality from above holds with $p^\ell = 0$ if $\lambda_2 = \sigma_2 = 0$ and the fourth with $p^g = 0$ if $\lambda_3 = \sigma_3 = 0$.

Therefore, in what follows we assume that $\lambda_2, \sigma_2, \lambda_3, \sigma_3 > 0$. In order to lower bound the three terms from above, we arrange the summands by using the abbreviations introduced in (3.10) as follows. For $T_{1k}^c + T_{2k}^c$, we have

$$T_{1k}^c(x, y) + T_{2k}^c(x, y)$$

$$= -\lambda_1 \left(x - y_\alpha(\rho_{Y,t})\right)_k \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma_1^2}{2} \left(x - y_\alpha(\rho_{Y,t})\right)_k^2 \delta_{x_k x_k}^{2,*} \phi_r(x, y) \tag{3.15a}$$

$$- \lambda_1 \left(x - y_\alpha(\rho_{Y,t})\right)_k \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_1^2}{2} \left(x - y_\alpha(\rho_{Y,t})\right)_k^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y), \tag{3.15b}$$

for $T_{1k}^\ell + T_{2k}^\ell$ we have

$$T_{1k}^\ell(x, y) + T_{2k}^\ell(x, y)$$

$$= -\lambda_2 (x - y)_k \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma_2^2}{2} (x - y)_k^2 \delta_{x_k x_k}^{2,*} \phi_r(x, y) \tag{3.16a}$$

$$- \lambda_2 (x - y)_k \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_2^2}{2} (x - y)_k^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y) \tag{3.16b}$$

and for $T_{1k}^g + T_{2k}^g$ we have

$$T_{1k}^g(x, y) + T_{2k}^g(x, y)$$

$$= -\lambda_3 \partial_{x_k}\mathcal{E}(x) \delta_{x_k}^* \phi_r(x, y) + \frac{\sigma_3^2}{2} (\partial_{x_k}\mathcal{E}(x))^2 \delta_{x_k x_k}^{2,*} \phi_r(x, y) \tag{3.17a}$$

$$- \lambda_3 \partial_{x_k}\mathcal{E}(x) \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_3^2}{2} (\partial_{x_k}\mathcal{E}(x))^2 \delta_{x_k x_k}^{2,Y} \phi_r(x, y). \tag{3.17b}$$

We now treat each of the two-part sums in (3.15a), (3.15b), (3.16a), (3.16b), (3.17a) and (3.17b) separately by employing a technique similar to the one used in the proof of [29, Proposition 2], which was developed originally to prove [28, Proposition 20].

Terms (3.15a), (3.16a) and (3.17a): Owed to their similar structure (in particular with respect to the denominator of the derivatives $\delta_{x_k}^* \phi_r$ and $\delta_{x_k x_k}^{2,*} \phi_r$), we can treat the three sums (3.15a), (3.16a) and (3.17a) simultaneously. Therefore, we consider the general formulation

$$-\lambda \Upsilon_k(x, y)\delta_{x_k}^* \phi_r(x, y) + \frac{\sigma^2}{2} \Upsilon_k^2(x, y)\delta_{x_k x_k}^{2,*} \phi_r(x, y) =: T_{1k}^*(x, y) + T_{2k}^*(x, y), \tag{3.18}$$

which matches (3.15a) when $\Upsilon_k(x, y) = (x - y_\alpha(\rho_{Y,t}))_k$, $\lambda = \lambda_1$ and $\sigma = \sigma_1$, (3.16a) when $\Upsilon_k(x, y) = (x - y)_k$, $\lambda = \lambda_2$ and $\sigma = \sigma_2$, and (3.17a) when $\Upsilon_k(x, y) = \partial_{x_k}\mathcal{E}(x)$, $\lambda = \lambda_3$ and $\sigma = \sigma_3$.

To achieve the desired lower bound over Ω_r, we introduce the subsets

$$K_{1k}^* := \left\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : |(x - x^*)_k| > \frac{\sqrt{c}}{2}r\right\}$$

and

$$K_{2k}^* := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : -\lambda \Upsilon_k(x, y) (x - x^*)_k \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \right.$$
$$\left. > \tilde{c} \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x - x^*)_k^2 \right\},$$

where $\tilde{c} := 2c - 1 \in (0, 1)$. For fixed k, we now decompose Ω_r according to

$$\Omega_r = \left((K_{1k}^*)^c \cap \Omega_r \right) \cup \left(K_{1k}^* \cap (K_{2k}^*)^c \cap \Omega_r \right) \cup \left(K_{1k}^* \cap K_{2k}^* \cap \Omega_r \right).$$

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

Subset $(K_{1k}^*)^c \cap \Omega_r$: We have $|(x - x^*)_k| \leq \frac{\sqrt{c}}{2} r$ for each $(x, y) \in (K_{1k}^*)^c$, which can be used to independently derive lower bounds for both summands in (3.18). For the first, we insert the expression for $\delta_{x_k}^* \phi_r(x, y)$ to get

$$T_{1k}^*(x, y) = \frac{r^2}{2} \lambda \Upsilon_k(x, y) \frac{(x - x^*)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \phi_r(x, y)$$
$$\geq -\frac{r^2}{2} \lambda \frac{|\Upsilon_k(x, y)||(x - x^*)_k|}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \phi_r(x, y) \geq -\frac{2\lambda C_\Upsilon \sqrt{c}}{(1 - c)^2 \frac{r}{2}} \phi_r(x, y) \qquad (3.19)$$
$$=: -p_1^{*,\Upsilon} \phi_r(x, y),$$

where, in the last inequality, we used that $(x, y) \in \Omega_r$, the definition of B and Assumption A3 to get the bound

$$|\Upsilon_k(x, y)| = \begin{cases} \left| (x - y_\alpha(\rho_{Y,t}))_k \right| \leq \frac{r}{2} + B, & \text{if } \Upsilon_k(x, y) = (x - y_\alpha(\rho_{Y,t}))_k, \\ |(x - y)_k| \leq \frac{r}{2}, & \text{if } \Upsilon_k(x, y) = (x - y)_k, \\ \left| \partial_{x_k} \mathcal{E}(x) \right| \leq \|\nabla \mathcal{E}(x)\|_2 \leq C_{\nabla \mathcal{E}} \|x - x^*\|_2 \\ \qquad \leq C_{\nabla \mathcal{E}} d \|x - x^*\|_\infty \leq C_{\nabla \mathcal{E}} d \frac{r}{2} & \text{if } \Upsilon_k(x, y) = \partial_{x_k} \mathcal{E}(x). \end{cases}$$
$$\leq \max \left\{ \frac{r}{2} + B, C_{\nabla \mathcal{E}} d \frac{r}{2} \right\} =: C_\Upsilon(r, B, d, C_{\nabla \mathcal{E}}). \qquad (3.20)$$

For the second summand, we insert the expression for $\delta_{x_k x_k}^{2,*} \phi_r(x, y)$ to obtain

$$T_{2k}^*(x, y) = \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \delta_{x_k x_k}^{2,*} \phi_r(x, y)$$
$$= \sigma^2 \left(\frac{r}{2} \right)^2 \Upsilon_k^2(x, y) \frac{2 \left(2 (x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2 - \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^4} \phi_r(x, y) \qquad (3.21)$$
$$\geq -\frac{\sigma^2 C_\Upsilon^2}{(1 - c)^4 \left(\frac{r}{2} \right)^2} \phi_r(x, y) =: -p_2^{*,\Upsilon} \phi_r(x, y),$$

where the last inequality uses $\Upsilon_k^2(x, y) \leq C_\Upsilon^2$.

Subset $K_{1k}^* \cap (K_{2k}^*)^c \cap \Omega_r$: As $(x, y) \in K_{1k}^*$, we have $|(x - x^*)_k| > \frac{\sqrt{c}}{2} r$. We observe that the sum in (3.18) is nonnegative for all (x, y) in this subset whenever

$$\left(-\lambda \Upsilon_k(x, y) (x - x^*)_k + \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \right) \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2$$
$$\leq \sigma^2 \Upsilon_k^2(x, y) \left(2 (x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2. \qquad (3.22)$$

The first term on the left-hand side in (3.22) can be bounded from above by exploiting that $v \in (K_{2k}^*)^c$ and by using the relation $\tilde{c} = 2c - 1$. More precisely, we have

$$
-\lambda \Upsilon_k(x, y)\, (x - x^*)_k \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \leq \tilde{c} \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y)\, (x - x^*)_k^2
$$

$$
= (2c - 1) \left(\frac{r}{2} \right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y)\, (x - x^*)_k^2 \leq \left(2\, (x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) \frac{\sigma^2}{2} \Upsilon_k^2(x, y)\, (x - x^*)_k^2,
$$

where the last inequality follows since $v \in K_{1k}^*$. For the second term on the left-hand side in (3.22), we can use $(1 - c)^2 \leq (2c - 1)c$ as per assumption, to get

$$
\frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y)(1 - c)^2 \left(\frac{r}{2} \right)^4
$$

$$
\leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y)(2c - 1) \left(\frac{r}{2} \right)^2 c \left(\frac{r}{2} \right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(2\, (x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2.
$$

Hence, (3.22) holds and we have that (3.18) is uniformly nonnegative on this subset.

Subset $K_{1k}^* \cap K_{2k}^* \cap \Omega_r$: As $(x, y) \in K_{1k}^*$, we have $|(x - x^*)_k| > \frac{\sqrt{c}}{2} r$. To start with we note that the first summand of (3.18) vanishes whenever $\sigma^2 \Upsilon_k^2(x, y) = 0$, provided $\sigma > 0$, so nothing needs to be done if $\Upsilon_k(x, y) = 0$. Otherwise, if $\sigma^2 \Upsilon_k^2(x, y) > 0$, we exploit $(x, y) \in K_{2k}^*$ to get

$$
\frac{\Upsilon_k(x, y)\, (x - x^*)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \geq \frac{-\,|\Upsilon_k(x, y)|\,|(x - x^*)_k|}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2}
$$

$$
> \frac{2\lambda \Upsilon_k(x, y)\, (x - x^*)_k}{\tilde{c} \left(\frac{r}{2} \right)^2 \sigma^2\, |\Upsilon_k(x, y)|\,|(x - x^*)_k|} \geq -\frac{8\lambda}{\tilde{c} r^2 \sigma^2}.
$$

Using this, the first summand of (3.18) can be bounded from below by

$$
T_{1k}^*(x, y) = \lambda \frac{r^2}{2} \frac{\Upsilon_k(x, y)\, (x - x^*)_k}{\left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2} \phi_r(x, y) \geq -\frac{4\lambda^2}{\tilde{c} \sigma^2} \phi_r(x, y) =: -p_3^{*,\Upsilon} \phi_r(x, y). \tag{3.23}
$$

For the second summand, the nonnegativity of $\sigma^2 \Upsilon_k^2(x, y)$ implies the nonnegativity, whenever

$$
2 \left(2\, (x - x^*)_k^2 - \left(\frac{r}{2} \right)^2 \right) (x - x^*)_k^2 \geq \left(\left(\frac{r}{2} \right)^2 - (x - x^*)_k^2 \right)^2.
$$

This holds for $v \in K_{1k}^*$, if $2(2c - 1)c \geq (1 - c)^2$ as implied by the assumption.

Term (3.16b): Recall that this term has the structure

$$
-\lambda_2\, (x - y)_k\, \delta_{x_k}^Y \phi_r(x, y) + \frac{\sigma_2^2}{2}\, (x - y)_k^2\, \delta_{x_k x_k}^{2, Y} \phi_r(x, y) =: T_{1k}^{Y,1}(x, y) + T_{2k}^{Y,1}(x, y). \tag{3.24}
$$

We first note that the first summand of (3.24) is always nonnegative since

$$
T_{1k}^{Y,1}(x, y) = \lambda_2 \frac{r^2}{2} \frac{(x - y)_k^2}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^2} \phi_r(x, y) \geq 0. \tag{3.25}
$$

For the second summand of (3.24), a direct computation shows

$$
T_{2k}^{Y,1}(x, y) = \sigma_2^2 \left(\frac{r}{2} \right)^2 (x - y)_k^2 \frac{3\, (x - y)_k^4 - \left(\frac{r}{2} \right)^4}{\left(\left(\frac{r}{2} \right)^2 - (x - y)_k^2 \right)^4} \phi_r(x, y),
$$

which is nonnegative on the set

$$
K_k^Y := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \left| (x - y)_k \right| > \frac{\sqrt{c}}{2} r \right\}
$$

for any $c \geq 1/\sqrt{3}$, as ensured by $(1-c)^2 \leq (2c-1)c$. On the complement $(K_k^Y)^c$, we have $\left|(x-y)_k\right| \leq \frac{\sqrt{c}}{2} r$, which can be used to bound

$$
\begin{aligned}
T_{2k}^{Y,1}(x,y) &= \sigma_2^2 \left(\frac{r}{2}\right)^2 (x-y)_k^2 \frac{3(x-y)_k^4 - \left(\frac{r}{2}\right)^4}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^4} \phi_r(x,y) \\
&\geq -\frac{\sigma_2^2 c}{(1-c)^4} \phi_r(x,y) =: -p^{Y,\Upsilon_\ell} \phi_r(x,y).
\end{aligned}
\tag{3.26}
$$

Terms (3.15b) and (3.17b): The final two terms to be controlled have again a similar structure of the form

$$
-\lambda \Upsilon_k(x,y) \delta_{x_k}^Y \phi_r(x,y) + \frac{\sigma^2}{2} \Upsilon_k^2(x,y) \delta_{x_k x_k}^{2,Y} \phi_r(x,y) =: T_{1k}^{Y,2}(x,y) + T_{2k}^{Y,2}(x,y),
\tag{3.27}
$$

where we recycle the notation introduced after (3.18), i.e., $\Upsilon_k(x,y) = (x - y_\alpha(\rho_{Y,t}))_k$, $\lambda = \lambda_1$ and $\sigma = \sigma_1$ in the case of (3.15b) and $\Upsilon_k(x,y) = \partial_{x_k} \mathcal{E}(x)$, $\lambda = \lambda_3$ and $\sigma = \sigma_3$ in the case of (3.17b).

The procedure for deriving lower bounds is similar to the one at the beginning with the exception that the denominator of the derivatives $\delta_{x_k}^Y \phi_r$ and $\delta_{x_k x_k}^{2,Y} \phi_r$ requires to introduce an adapted decomposition of Ω_r. To be more specific, we define the subsets

$$
K_{1k}^Y := \left\{ (x,y) \in \mathbb{R}^d \times \mathbb{R}^d : \left|(x-y)_k\right| > \frac{\sqrt{c}}{2} r \right\}
$$

and

$$
\begin{aligned}
K_{2k}^Y := \Big\{ (x,y) \in \mathbb{R}^d \times \mathbb{R}^d : &-\lambda \Upsilon_k(x,y)(x-y)_k \left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2 \\
&> \tilde{c} \left(\frac{r}{2}\right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x,y)(x-y)_k^2 \Big\},
\end{aligned}
$$

where $\tilde{c} := 2c - 1 \in (0,1)$. For fixed k, we now decompose Ω_r according to

$$
\Omega_r = \left((K_{1k}^Y)^c \cap \Omega_r\right) \cup \left(K_{1k}^Y \cap (K_{2k}^Y)^c \cap \Omega_r\right) \cup \left(K_{1k}^Y \cap K_{2k}^Y \cap \Omega_r\right).
$$

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

Subset $(K_{1k}^Y)^c \cap \Omega_r$: We have $\left|(x-y)_k\right| \leq \frac{\sqrt{c}}{2} r$ for each $(x,y) \in (K_{1k}^Y)^c$, which can be used to independently derive lower bounds for both summands in (3.27). For the first summand, we insert the expression for $\delta_{x_k}^Y \phi_r(x,y)$ to get

$$
\begin{aligned}
T_{1k}^{Y,2}(x,y) &= \frac{r^2}{2} \lambda \Upsilon_k(x,y) \frac{(x-y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \phi_r(x,y) \\
&\geq -\frac{r^2}{2} \lambda \frac{\left|\Upsilon_k(x,y)\right|\left|(x-y)_k\right|}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \phi_r(x,y) \geq -\frac{2\lambda C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} \phi_r(x,y) \\
&=: -p_1^{Y,\Upsilon} \phi_r(x,y),
\end{aligned}
\tag{3.28}
$$

where we recall from above that $\Upsilon_k(x,y) \leq C_\Upsilon$, which was used in the last inequality. For the second summand, we insert the expression for $\delta_{x_k x_k}^{2,Y} \phi_r(x,y)$ to obtain

$$
\begin{aligned}
T_{2k}^{Y,2}(x,y) &= \sigma^2 \left(\frac{r}{2}\right)^2 \Upsilon_k^2(x,y) \frac{2\left(2(x-y)_k^2 - \left(\frac{r}{2}\right)^2\right)(x-y)_k^2 - \left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^4} \phi_r(x,y) \\
&\geq -\frac{\sigma^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} \phi_r(x,y) =: -p_2^{Y,\Upsilon} \phi_r(x,y),
\end{aligned}
\tag{3.29}
$$

where the last inequality uses $\Upsilon_k^2(x,y) \leq C_\Upsilon^2$.

Subset $K_{1k}^Y \cap (K_{2k}^Y)^c \cap \Omega_r$: As $(x, y) \in K_{1k}^Y$, we have $|(x-y)_k| > \frac{\sqrt{c}}{2}r$. We observe that the sum in (3.27) is nonnegative for all (x, y) in this subset whenever

$$\left(-\lambda \Upsilon_k(x, y) (x-y)_k + \frac{\sigma^2}{2} \Upsilon_k^2(x, y)\right) \left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2$$
$$\leq \sigma^2 \Upsilon_k^2(x, y) \left(2 (x-y)_k^2 - \left(\frac{r}{2}\right)^2\right) (x-y)_k^2. \tag{3.30}$$

The first term on the left-hand side in (3.30) can be bounded from above exploiting that $v \in (K_{2k}^Y)^c$ and by using the relation $\tilde{c} = 2c - 1$. More precisely, we have

$$-\lambda \Upsilon_k(x, y) (x-y)_k \left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2 \leq \tilde{c} \left(\frac{r}{2}\right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x-y)_k^2$$
$$= (2c - 1) \left(\frac{r}{2}\right)^2 \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x-y)_k^2 \leq \left(2 (x-y)_k^2 - \left(\frac{r}{2}\right)^2\right) \frac{\sigma^2}{2} \Upsilon_k^2(x, y) (x-y)_k^2,$$

where the last inequality follows since $v \in K_{1k}^Y$. For the second term on the left-hand side in (3.30), we can use $(1 - c)^2 \leq (2c - 1)c$ as per assumption, to get

$$\frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y)(1 - c)^2 \left(\frac{r}{2}\right)^4$$
$$\leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y)(2c - 1) \left(\frac{r}{2}\right)^2 c \left(\frac{r}{2}\right)^2 \leq \frac{\sigma^2}{2} \Upsilon_k^2(x, y) \left(2 (x-y)_k^2 - \left(\frac{r}{2}\right)^2\right) (x-y)_k^2.$$

Hence, (3.30) holds and we have that (3.27) is uniformly nonnegative on this subset.

Subset $K_{1k}^Y \cap K_{2k}^Y \cap \Omega_r$: As $(x, y) \in K_{1k}^Y$, we have $|(x-y)_k| > \frac{\sqrt{c}}{2}r$. To start with we note that the first summand of (3.27) vanishes whenever $\sigma^2 \Upsilon_k^2(x, y) = 0$, provided $\sigma > 0$, so nothing needs to be done if $\Upsilon_k(x, y) = 0$. Otherwise, if $\sigma^2 \Upsilon_k^2(x, y) > 0$, we exploit $(x, y) \in K_{2k}^Y$ to get

$$\frac{\Upsilon_k(x, y) (x-y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \geq \frac{-|\Upsilon_k(x, y)| |(x-y)_k|}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2}$$
$$> \frac{2\lambda \Upsilon_k(x, y) (x-y)_k}{\tilde{c} \left(\frac{r}{2}\right)^2 \sigma^2 |\Upsilon_k(x, y)| |(x-y)_k|} \geq -\frac{8\lambda}{\tilde{c}r^2\sigma^2}.$$

Using this, the first summand of (3.27) can be bounded from below by

$$T_{1k}^{Y,2}(x, y) = \lambda \frac{r^2}{2} \frac{\Upsilon_k(x, y) (x-y)_k}{\left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2} \phi_r(x, y) \geq -\frac{4\lambda^2}{\tilde{c}\sigma^2} \phi_r(x, y) =: -p_3^{Y,\Upsilon} \phi_r(x, y). \tag{3.31}$$

For the second summand, the nonnegativity of $\sigma^2 \Upsilon_k^2(x, y)$ implies the nonnegativity, whenever

$$2 \left(2 (x-y)_k^2 - \left(\frac{r}{2}\right)^2\right) (x-y)_k^2 \geq \left(\left(\frac{r}{2}\right)^2 - (x-y)_k^2\right)^2.$$

This holds for $v \in K_{1k}^Y$, if $2(2c - 1)c \geq (1 - c)^2$ as implied by the assumption.

Concluding the proof: Combining the formerly established lower bounds (3.19), (3.21), (3.23), (3.25), (3.26), (3.28), (3.29) and (3.31), we obtain for the constants p^c, p^ℓ and p^g defined at the beginning of the proof

$$p^c = p_1^{*,\Upsilon_c} + p_2^{*,\Upsilon_c} + p_3^{*,\Upsilon_c} + p_1^{Y,\Upsilon_c} + p_2^{Y,\Upsilon_c} + p_3^{Y,\Upsilon_c} = 2\left(\frac{2\lambda_1 C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_1^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_1^2}{\tilde{c}\sigma_1^2} \right)$$

$$p^\ell = p_1^{*,\Upsilon_\ell} + p_2^{*,\Upsilon_\ell} + p_3^{*,\Upsilon_\ell} + p^{Y,\Upsilon_\ell} = \frac{2\lambda_2 C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_2^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_2^2}{\tilde{c}\sigma_2^2} + \frac{\sigma_2^2 c}{(1-c)^4} \qquad (3.32)$$

$$p^g = p_1^{*,\Upsilon_g} + p_2^{*,\Upsilon_g} + p_3^{*,\Upsilon_g} + p_1^{Y,\Upsilon_g} + p_2^{Y,\Upsilon_g} + p_3^{Y,\Upsilon_g} = 2\left(\frac{2\lambda_3 C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_3^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_3^2}{\tilde{c}\sigma_3^2} \right).$$

Together with (3.14) and by using the evolution of ϕ_r as in (3.13), we eventually obtain

$$\frac{d}{dt} \iint \phi_r \, d\rho_t \geq -d \left(p^c + p^\ell + p^g \right) \iint \phi_r \, d\rho_t$$

$$\geq -d \sum_{i=1}^{3} \omega_i \left((1 + \mathbb{1}_{i \neq 2}) \left(\frac{2\lambda_i C_\Upsilon \sqrt{c}}{(1-c)^2 \frac{r}{2}} + \frac{\sigma_i^2 C_\Upsilon^2}{(1-c)^4 \left(\frac{r}{2}\right)^2} + \frac{4\lambda_i^2}{\tilde{c}\sigma_i^2} \right) + \mathbb{1}_{i=2} \frac{\sigma_2^2 c}{(1-c)^4} \right) \iint \phi_r \, d\rho_t$$

$$= -q \iint \phi_r \, d\rho_t,$$

where q is defined implicitly and where $\omega_i = \mathbb{1}_{\lambda_i > 0}$ for $i \in \{1, 2, 3\}$. Notice that $\omega_1 = 1$ since $\lambda_1 > 0$ by assumption. An application of Grönwall's inequality concludes the proof. □

3.4. Proof of Theorem 2.5

We now have all necessary tools at hand to prove the global mean-field law convergence result for CBO with memory effects and gradient information by rigorously combining the formerly discussed statements.

Proof of Theorem 2.5. If $\mathcal{V}(\rho_0) = 0$, there is nothing to be shown since in this case $\rho_0 = \delta_{(x^*, x^*)}$. Thus, let $\mathcal{V}(\rho_0) > 0$ in what follows.

Let us first choose the parameter α such that

$$\alpha > \alpha_0 := \frac{1}{q_\varepsilon} \left(\log\left(\frac{2^{d+2}\sqrt{d}}{c\,(\vartheta, \chi_1, \lambda_1, \sigma_1)} \right) + \max\left\{ \frac{1}{2}, \frac{p}{(1-\vartheta)\chi_1} \right\} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon} \right) - \log \rho_0\left(\Omega_{r_\varepsilon/2}\right) \right), \tag{3.33}$$

where we introduce the definitions

$$c\,(\vartheta, \chi_1, \lambda_1, \sigma_1) := \min\left\{ \frac{\vartheta}{2} \frac{\chi_1}{2\sqrt{2}\left(\lambda_1 + \sigma_1^2\right)}, \sqrt{\frac{\vartheta}{2}\frac{\chi_1}{\sigma_1^2}} \right\} \tag{3.34}$$

as well as

$$q_\varepsilon := \frac{1}{2} \min\left\{ \left(\eta \frac{c\,(\vartheta, \chi_1, \lambda_1, \sigma_1)\sqrt{\varepsilon}}{2\sqrt{d}} \right)^{1/\nu}, \mathcal{E}_\infty \right\} \quad \text{and} \quad r_\varepsilon := \max_{s \in [0, R_0]} \left\{ \max_{v \in B_s^\infty(x^*)} \mathcal{E}(v) \leq q_\varepsilon \right\}. \tag{3.35}$$

Moreover, p is as given in (3.12) in Proposition 3.8 with $B = c\,(\vartheta, \chi_1, \lambda_1, \sigma_1)\sqrt{\mathcal{V}(\rho_0)}$ in C_Υ and with $r = r_\varepsilon$. By construction, $q_\varepsilon > 0$ and $r_\varepsilon \leq R_0$. Furthermore, recalling the notation $\mathcal{E}_r = \sup_{v \in B_r^\infty(x^*)} \mathcal{E}(v)$ from Proposition 3.6, we have $q_\varepsilon + \mathcal{E}_{r_\varepsilon} \leq 2q_\varepsilon \leq \mathcal{E}_\infty$ according to the definition of r_ε. Since $q_\varepsilon > 0$, the continuity of $\mathcal{E}$ ensures that there exists $s_{q_\varepsilon} > 0$ such that $\mathcal{E}(v) \leq q_\varepsilon$ for all $v \in B_{s_{q_\varepsilon}}^\infty(x^*)$, yielding also $r_\varepsilon > 0$.

Let us now define the time horizon $T_\alpha \geq 0$ by

$$T_\alpha := \sup\left\{ t \geq 0 : \mathcal{V}(\rho_{t'}) > \varepsilon \text{ and } \left\| y_\alpha(\rho_{Y,t'}) - x^* \right\|_2 < C(t') \text{ for all } t' \in [0, t] \right\} \tag{3.36}$$

with $C(t) := c\,(\vartheta, \chi_1, \lambda_1, \sigma_1)\sqrt{\mathcal{V}(\rho_t)}$. Notice for later use that $C(0) = B$.

Our aim now is to show that $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in \left[\frac{(1-\vartheta)\chi_1}{(1+\vartheta/2)\chi_2} T^*, T^*\right]$ and that we have at least exponential decay of $\mathcal{V}(\rho_t)$ until time T_α, i.e., until the accuracy ε is reached.

First, however, we verify that $T_\alpha > 0$, which is due to the continuity of $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2$ since $\mathcal{V}(\rho_0) > \varepsilon$ and $\left\|y_\alpha(\rho_{Y,0}) - x^*\right\|_2 < C(0)$ at time 0. While the former is a consequence of the assumption, the latter follows from Proposition 3.6 with q_ε and r_ε as defined in (3.35), which allows to show that

$$
\begin{aligned}
\left\|y_\alpha(\rho_{Y,0}) - x^*\right\|_2 &\le \frac{\sqrt{d}\left(q_\varepsilon + \mathcal{E}_{r_\varepsilon}\right)^\nu}{\eta} + \frac{\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,0}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \int \|y - x^*\|_2 \, d\rho_{Y,0}(y) \\
&\le \frac{\sqrt{d}\left(q_\varepsilon + \mathcal{E}_{r_\varepsilon}\right)^\nu}{\eta} + \frac{\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,0}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \iint \|y - x\|_2 + \|x - x^*\|_2 \, d\rho_0(x,y) \\
&\le \frac{c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\varepsilon}}{2} + \frac{2\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,0}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \sqrt{\mathcal{V}(\rho_0)} \\
&\le c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\varepsilon} < c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\mathcal{V}(\rho_0)} = C(0).
\end{aligned}
$$

The first inequality in the last line holds by the choice of α in (3.33) and by noticing that $\Omega_{r_\varepsilon/2} \subset \mathbb{R}^d \times B_{r_\varepsilon}^\infty(x^*)$ and thus $\rho_0(\Omega_{r_\varepsilon/2}) \le \rho_{Y,0}\left(B_{r_\varepsilon}^\infty(x^*)\right)$.

Next, we show that the functional $\mathcal{V}(\rho_t)$ is sandwiched between two exponentially decaying functions with rates $(1 - \vartheta)\chi_1$ and $(1 + \vartheta/2)\chi_2$, respectively. More precisely, we prove that, up to time T_α, $\mathcal{V}(\rho_t)$ decays

  (i) at least exponentially fast (with rate $(1 - \vartheta)\chi_1$), and

  (ii) at most exponentially fast (with rate $(1 + \vartheta/2)\chi_2$).

To obtain (i), recall that Corollary 3.3 provides an upper bound on the time derivative of $\mathcal{V}(\rho_t)$ given by

$$
\frac{d}{dt}\mathcal{V}(\rho_t) \le -\chi_1 \mathcal{V}(\rho_t) + 2\sqrt{2}\left(\lambda_1 + \sigma_1^2\right)\sqrt{\mathcal{V}(\rho_t)}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2 + \sigma_1^2\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2^2 \tag{3.37}
$$

with χ_1 as in (2.12a) being strictly positive by assumption. By combining (3.37) and the definition of T_α in (3.36), we have by construction

$$
\frac{d}{dt}\mathcal{V}(\rho_t) \le -(1 - \vartheta)\chi_1 \mathcal{V}(\rho_t) \quad \text{for all } t \in (0, T_\alpha).
$$

Analogously, for (ii), by Corollary 3.5, we obtain a lower bound on the time derivative of $\mathcal{V}(\rho_t)$ given by

$$
\begin{aligned}
\frac{d}{dt}\mathcal{V}(\rho_t) &\ge -\chi_2 \mathcal{V}(\rho_t) - 2\sqrt{2}\left(\lambda_1 + \sigma_1^2\right)\sqrt{\mathcal{V}(\rho_t)}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2 \\
&\ge -(1 + \vartheta/2)\chi_2 \mathcal{V}(\rho_t) \quad \text{for all } t \in (0, T_\alpha),
\end{aligned} \tag{3.38}
$$

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

$$
\mathcal{V}(\rho_t) \le \mathcal{V}(\rho_0)\exp\left(-(1 - \vartheta)\chi_1 t\right), \tag{3.39a}
$$

$$
\mathcal{V}(\rho_t) \ge \mathcal{V}(\rho_0)\exp\left(-(1 + \vartheta/2)\chi_2 t\right), \tag{3.39b}
$$

thereby proving (i) and (ii). The definition of T_α together with the one of $C(t)$ permits to control

$$
\max_{t \in [0, T_\alpha]}\left\|y_\alpha(\rho_{Y,t}) - x^*\right\|_2 \le \max_{t \in [0, T_\alpha]} C(t) \le C(0). \tag{3.40}
$$

To conclude, it remains to prove $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ with $T_\alpha \in \left[\frac{(1-\vartheta)\chi_1}{(1+\vartheta/2)\chi_2} T^*, T^*\right]$. To this end, we consider the following three cases separately.

Case $T_\alpha \ge T^*$: If $T_\alpha \ge T^*$, the time-evolution bound of $\mathcal{V}(\rho_t)$ from (3.39a) combined with the definition of T^* in (2.13) allows to immediately infer $\mathcal{V}(\rho_{T^*}) \le \varepsilon$. Therefore, with $\mathcal{V}(\rho_t)$ being continuous, $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$ and $T_\alpha = T^*$ according to the definition of T_α in (3.36).

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) \leq \varepsilon$: By continuity of $\mathcal{V}(\rho_t)$, it holds for T_α as defined in (3.36), $\mathcal{V}(\rho_{T_\alpha}) = \varepsilon$. Thus, $\varepsilon = \mathcal{V}(\rho_{T_\alpha}) \geq \mathcal{V}(\rho_0) \exp\left(-(1+\vartheta/2)\chi_2 T_\alpha\right)$ as a consequence of the time-evolution bound (3.39b). The latter can be reordered as

$$\frac{(1-\vartheta)\chi_1}{(1+\vartheta/2)\chi_2} T^* = \frac{1}{(1+\vartheta/2)\chi_2} \log\left(\frac{\mathcal{V}(\rho_0)}{\varepsilon}\right) \leq T_\alpha < T^*.$$

Case $T_\alpha < T^*$ and $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$: We will prove that this case can actually not occur by showing that $\left\| y_\alpha(\rho_{Y,T_\alpha}) - x^* \right\|_2 < C(T_\alpha)$ for the α chosen in (3.33). In fact, if both $\mathcal{V}(\rho_{T_\alpha}) > \varepsilon$ and $\left\| y_\alpha(\rho_{Y,T_\alpha}) - x^* \right\|_2 < C(T_\alpha)$ held true simultaneously, this would contradict the definition of T_α in (3.36). To obtain this contradiction, we apply again Proposition 3.6 with q_ε and r_ε as before to get

$$
\begin{aligned}
\left\| y_\alpha(\rho_{Y,T_\alpha}) - x^* \right\|_2 &\leq \frac{\sqrt{d}\left(q_\varepsilon + \mathcal{E}_{r_\varepsilon}\right)^\nu}{\eta} + \frac{\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,T_\alpha}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \int \|y - x^*\|_2 \, d\rho_{Y,T_\alpha}(y) \\
&\leq \frac{\sqrt{d}\left(q_\varepsilon + \mathcal{E}_{r_\varepsilon}\right)^\nu}{\eta} + \frac{\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,T_\alpha}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \iint \|y - x\|_2 + \|x - x^*\|_2 \, d\rho_{T_\alpha}(x,y) \\
&\leq \frac{c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\varepsilon}}{2} + \frac{2\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,T_\alpha}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \sqrt{\mathcal{V}(\rho_{T_\alpha})} \\
&< \frac{c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_{Y,T_\alpha}\left(B_{r_\varepsilon}^\infty(x^*)\right)} \sqrt{\mathcal{V}(\rho_{T_\alpha})}.
\end{aligned}
\tag{3.41}
$$

Since, thanks to (3.40), we have $\max_{t \in [0, T_\alpha]} \|y_\alpha(\rho_{Y,t}) - x^*\|_2 \leq B$ for $B = C(0)$, which in particular does not depend on α, Proposition 3.8 guarantees the existence of $p > 0$ independent of α (but dependent on B and r_ε) with

$$
\begin{aligned}
\rho_{Y,T_\alpha}(B_{r_\varepsilon}^\infty(x^*)) &\geq \left(\iint \phi_{r_\varepsilon}(x,y)\, d\rho_0(x,y)\right) \exp\left(-pT_\alpha\right) \\
&\geq \frac{1}{2^d}\, \rho_0\left(\Omega_{r_\varepsilon/2}\right) \exp\left(-pT^*\right) > 0.
\end{aligned}
$$

Here we use that $(x^*, x^*) \in \mathrm{supp}\,(\rho_0)$ to bound the initial mass ρ_0 and the fact that ϕ_r from Lemma 3.7 is bounded from below on $\Omega_{r/2}$ by $1/2^d$. With this, we can continue the chain of inequalities in (3.41)3 to obtain

$$
\begin{aligned}
\left\| y_\alpha(\rho_{Y,T_\alpha}) - x^* \right\|_2 &< \frac{c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\mathcal{V}(\rho_{T_\alpha})}}{2} + \frac{2^{d+1}\sqrt{d}\exp\left(-\alpha q_\varepsilon\right)}{\rho_0\left(\Omega_{r_\varepsilon/2}\right)\exp\left(-pT^*\right)} \sqrt{\mathcal{V}(\rho_{T_\alpha})} \\
&\leq c\left(\vartheta, \chi_1, \lambda_1, \sigma_1\right)\sqrt{\mathcal{V}(\rho_{T_\alpha})} = C(T_\alpha),
\end{aligned}
$$

with the first inequality in the last line holding due to the choice of α in (3.33). This gives the desired contradiction, again thanks to the continuity of $t \mapsto \mathcal{V}(\rho_t)$ and $t \mapsto \left\| y_\alpha(\rho_{Y,t}) - x^* \right\|_2$. $\square$

4. Numerical experiments

In the first part of this section, we comment on how to efficiently implement a numerical scheme for the CBO dynamics (1.1) which allows to integrate memory mechanisms without additional computational complexity. Afterwards, we numerically demonstrate the benefit of memory effects and gradient information at the example of interesting real-world inspired applications.

4.1. Implementational aspects

Discretising the interacting particle system (1.1) in time by means of the Euler-Maruyama method [37] with prescribed time step size Δt results in the implementable numerical scheme

$$
\begin{aligned}
X_{k+1}^i = {} & X_k^i - \Delta t \lambda_1\big(X_k^i - y_\alpha(\widehat{\rho}_{Y,k}^N)\big) - \Delta t \lambda_2\big(X_k^i - Y_k^i\big) - \Delta t \lambda_3 \nabla\mathcal{E}(X_k^i) \\
& + \sigma_1 D\big(X_k^i - y_\alpha(\widehat{\rho}_{Y,k}^N)\big) B_k^{1,i} + \sigma_2 D\big(X_k^i - Y_k^i\big) B_k^{2,i} + \sigma_3 D\big(\nabla\mathcal{E}(X_k^i)\big) B_k^{3,i},
\end{aligned} \tag{4.1a}
$$

$$
Y_{k+1}^i = Y_k^i + \Delta t \kappa \, \big(X_{k+1}^i - Y_k^i\big) S^{\beta,\theta}\big(X_{k+1}^i, Y_k^i\big), \tag{4.1b}
$$

where $((B_k^{m,i})_{k=0,\dots,K-1})_{i=1,\dots,N}$ are independent, identically distributed Gaussian random vectors in $\mathbb{R}^d$ with zero mean and covariance matrix $\Delta t \mathrm{Id}$ for $m \in \{1, 2, 3\}$.

We notice that, compared to standard CBO, see, e.g., [28, Equation (2)], the way the historical best position is updated in (4.1b) (recall the definition of $S^{\beta,\theta}$ from equation (1.4)) requires one additional evaluation of the objective function per particle in each time step, which raises the computational complexity of the numerical scheme substantially if computing $\mathcal{E}$ is costly and the dominating part. However, for the parameter choices $\kappa = 1/\Delta t$, $\theta = 0$ and $\beta = \infty$, in place of (4.1b), we obtain the update rule

$$
Y_{k+1}^i = \begin{cases} X_{k+1}^i, & \text{if } \mathcal{E}(X_{k+1}^i) < \mathcal{E}(Y_k^i), \\ Y_k^i, & \text{else,} \end{cases} \tag{4.2}
$$

which is how one expects a memory mechanism to be implemented. This way allows to recycle in time step k the computations made in the previous step and thus leads to no additional computational cost as consequence of using memory effects. The memory consumption, on the other hand, is approximately twice as high as in standard CBO.

4.2. A benchmark problem in optimisation: the Rastrigin function

Let us validate in this section the numerical observation made in Figure 2a in the introduction about the benefit of memory effects. Namely, it has been observed in several prior works that a higher noise level can enhance the success of CBO. To rule out that the improved performance for $\lambda_2 > 0$ in Figure 2a originates solely from the larger present noise as consequence of the additional noise term associated with the memory drift, we replicate in Figure 3 the experiments with the exception of setting $\sigma_2 = 0$. The obtained results confirm that already the usage of memory effects together with a memory drift improves the performance. However, we also notice that an additional noise term further increases the success probability.

4.3. A machine learning example

As a first real-world inspired application, we now investigate the influence of memory mechanisms in a high-dimensional benchmark problem in machine learning, which is well-understood in the literature, namely the training of a shallow and a convolutional NN (CNN) classifier for the MNIST dataset of handwritten digits [47].

The experimental setting is the one of [29, Section 4] with tested architectures as described in Figure 4. While it is not our aim to challenge the state of the art at this task by employing very sophisticated architectures, we demonstrate that CBO is on par with stochastic gradient descent without requiring time-consuming hyperparameter tuning.

To train the learnable parameters θ of the NNs, we minimise the empirical risk $\mathcal{E}(\theta) = \frac{1}{M} \sum_{j=1}^{M} \ell(f_\theta(x^j), y^j)$, where f_θ denotes the forward pass of the NN and $(x^j, y^j)_{j=1}^{M}$ the M training samples consisting of image and label. As loss ℓ we choose the categorical crossentropy loss $\ell(\widehat{y}, y) = -\sum_{k=0}^{9} y_k \log(\widehat{y}_k)$ with $\widehat{y} = f_\theta(x)$ denoting the output of the NN for a sample (x, y).

Our implementation is the one of [29, Section 4], which includes concepts from [14] and [26, Section 2.2]. Firstly, mini-batching is employed when evaluating $\mathcal{E}$ and when computing the consensus point y_α,
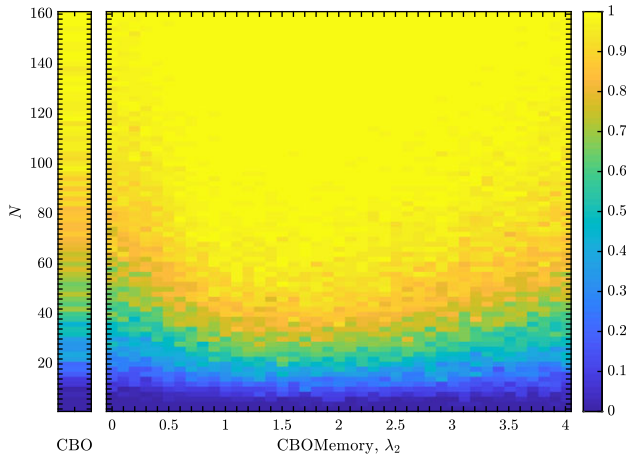
Figure 3. *Success probability of CBO without (left separate column) and with memory effects for different values of the parameter $\lambda_2 \in [0, 4]$ (right phase diagram) when optimising the Rastrigin function in dimension $d = 4$ in the setting of Figure 2a with the exception of setting $\sigma_2 = 0$. In this way we validate that the presence of memory effects is responsible for the improved performance and not just a higher noise level.*
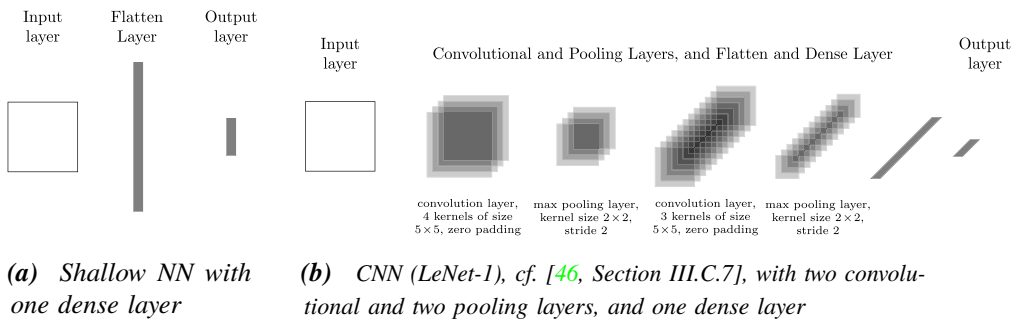

(a) *Shallow NN with one dense layer*

(b) *CNN (LeNet-1), cf. [46, Section III.C.7], with two convolutional and two pooling layers, and one dense layer*

Figure 4. *NN architectures used in the experiments of Section 4.3. Images are represented as 28×28 matrices with entries in $[0, 1]$. For the shallow NN in (a) the input is reshaped into a vector $x \in \mathbb{R}^{728}$ which is then passed through a dense layer of the form $ReLU(Wx + b)$ with trainable weights $W \in \mathbb{R}^{10 \times 728}$ and bias $b \in \mathbb{R}^{10}$. The learnable parameters of the CNN in (b) are the kernels and the final dense layer. Both networks include a batch normalisation step after each ReLU activation function and a softmax activation in the last layer in order to be able to interpret the output as a probability distribution over the digits. We denote the trainable parameters of the NN by θ. The shallow NN has 7850 and the CNN 2112. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Applications of Evolutionary Computation, Convergence of Anisotropic Consensus-Based Optimization in Mean-Field Law, M. Fornasier, T. Klock, K. Riedl, © 2022.)*

which means that $\mathcal{E}$ is evaluated on a random subset of size $n_\mathcal{E} = 60$ of the training dataset and y_α is computed from a random subset of size $n_N = 10$ of all $N = 100$ particles. Secondly, a cooling strategy for α and the noise parameters is used. More precisely, α is doubled each epoch, while σ_1 and σ_2 follow the schedule $\sigma_{i,\text{epoch}} = \sigma_{i,0}/\log_2(\text{epoch} + 2)$ for $i = 1, 2$.

In Figure 5, we report the testing accuracies and the training risks evaluated at the consensus point based on a random sample of the training set of size 10,000 for both the shallow NN and the CNN when trained with one of three algorithms: standard CBO without memory effects as obtained when
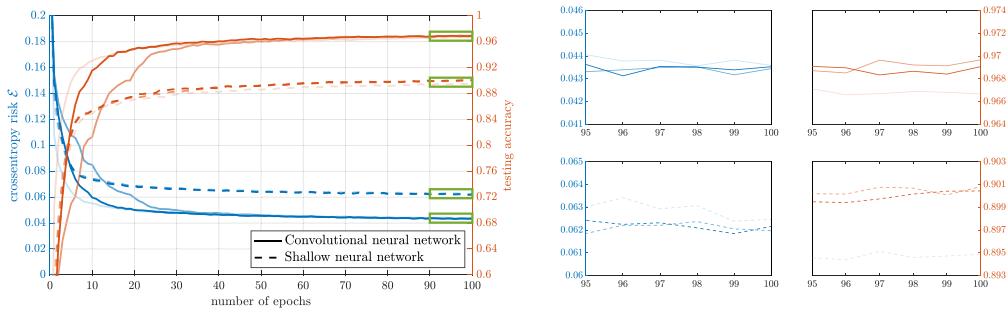
(a) *Testing accuracy and empirical risk plots for the shallow NN and the CNN when trained with CBO without memory effects (lightest lines), with memory effects but without memory drift (line with intermediate opacity) and with memory effects and memory drift (darkest lines)*

(b) *Zooming into the testing accuracies (right column) and the empirical risks (left column) during the final 5 epochs (highlighted with green boxes in (a)) fortheshallow NN (bottom row) and the CNN (top row)*

Figure 5. *Comparison of the performances (testing accuracy and training loss) of a shallow NN (dashed lines) and a CNN (solid lines) with architectures as described in Figure 4, when trained with CBO without memory effects (lightest lines), with memory effects but without memory drift (line with intermediate opacity) and with memory effects and memory drift (darkest lines). Depicted are the accuracies on a test dataset (orange lines) and the values of the objective function $\mathcal{E}$ evaluated on a random sample of the training set of size 10,000 (blue lines). We observe that memory effects slightly improve the final accuracies while slowing down the training process initially.*

discretising [29, Equation (1)], CBO with memory effects but without memory drift as in equation (4.1) with $\lambda_2 = \sigma_2 = 0$, and CBO with memory effects and memory drift as in equation (4.1) with $\lambda_2 = 0.4$ and $\sigma_2 = \lambda_2\sigma_1$. The remaining parameters are $\lambda_1 = 1$, $\sigma_{1,0} = \sqrt{0.4}$, $\alpha_{\text{initial}} = 50$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, and discrete time step size $\Delta t = 0.1$. We train for 100 epochs and use $N = 100$ particles, which are initialised according to $\mathcal{N}\big((0,\ldots,0)^T, \text{Id}\big)$. All results are averaged over 5 training runs.

As concluded already in [29, Section 4], we obtain accuracies comparable to SGD, cf. [46, Figure 9]. Moreover, we see slightly improved results when exploiting memory effects. However, we also notice that memory mechanisms slow down the training process initially.

4.4. A compressed sensing example

In the final numerical section of this paper, we showcase an application where gradient information turns out to be indispensable for the success of CBO methods, namely an experiment in compressed sensing [30], which has become a very active and profitable field of research since the seminal works [11, 22] about two decades ago.

One of the most common problems encountered in engineering and technology is concerned with the inference of information about an unknown signal $x^* \in \mathbb{R}^d$ from (linear) measurements $b \in \mathbb{R}^m$. While classical linear algebra suggests that the number of measurements m must be at least as large as the dimensionality d of the signal, in many applications measurements are costly, time-consuming or both, making it desirable to reduce their number to the absolute minimum. Very often one aims at $m \ll d$, since real-world signals usually live in high-dimensional spaces. In general, this would be an impossible task. However, in typical practical scenarios additional information about the quantity of interest x^* is available, which indeed allows to reconstruct signals from few measurements b. An empirically observed assumption about real-world signals is compressibility, meaning that they can be well-approximated by

sparse vectors, i.e., vectors whose components are for the most part zero. Exploiting sparsity enables us to solve the underdetermined system $Ax^* = b$ efficiently in both theory and practice. Compressed sensing is concerned with the task of designing a measurement process $A \in \mathbb{R}^{m \times d}$ together with a reconstruction algorithm capable of recovering the sparse solution x^* from the set of solutions consistent with the measurements. This can be formulated as the nonconvex combinatorial optimisation

$$\min \|x\|_0 \quad \text{subject to } Ax = b,$$

where $\|x\|_0$ is colloquially referred to as ℓ_0-'norm' and denotes the number of non-zero elements of x. Solving ℓ_0-minimisation is however NP-hard in general, which lead researchers to consider the convex relaxation

$$\min \|x\|_1 \quad \text{subject to } Ax = b. \tag{4.3}$$

ℓ_1-minimisation is easy to solve by means of established methods from convex optimisation and provably recovers the correct solution for a suitable measurement matrix A. The remaining question is about the correct way of inferring information about the signal through measurements. Remarkably and responsible for the wide success of compressed sensing is that random matrices enjoy properties such as the null space or restricted isometry property, which guarantee successful recovery, for $m \gtrsim s \log(d/s)$, where s denotes the sparsity of the signal x^*, i.e., $s = \|x^*\|_0$. Up to the logarithmic factor in the ambient dimension d, this is optimal, since in theory $m = 2s$ measurements are necessary and sufficient to reconstruct every s-sparse vector.

In the numerical experiments following, we resort to the regularised variant of the sparse recovery problem

$$\min \; \mathcal{E}(x) \quad \text{with } \mathcal{E}(x) = \frac{1}{2} \|Ax - b\|_2^2 + \mu \|x\|_p^p \tag{4.4}$$

for a suitable regularisation parameter $\mu > 0$. For $p = 1$ we obtain the regularisation of (4.3), whereas for $p < 1$ the optimisation (4.4) becomes nonconvex. Our results in Figures 2b and 6 show that CBO with gradient information is capable of solving the convex but also the nonconvex optimisation problem (4.4) with $p = 1/2$ with already very few measurements. As parameters of the CBO algorithm, which is obtained as a Euler-Maruyama discretisation of equation (1.1), we choose in both cases the time horizon $T = 20$, time step size $\Delta t = 0.01$, $\alpha = 100$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, $\lambda_1 = 1$, $\lambda_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_3 = 0$. We use either $N = 10$ or $N = 100$ particles, which is specified in the respective caption. After running the CBO algorithm, a post-processing step is performed, in which the support of the suspected sparse vector is identified by checking which entries are not smaller than 0.01 before the final sparse solution is then obtained by solving the linear system constrained to this support.

The depicted success probabilities are averaged over 100 runs of CBO. In Figure 2b, we solve the sparse recovery problem in the convex setting for an 8-sparse 200-dimensional signal with $p = 1$ using CBO without and with gradient information with merely 10 particles. We observe that gradient information is indispensable to be able to identify the correct sparse solution and standard CBO would fail in such task. In Figure 6, we conduct a slightly lower-dimensional experiment with a 2-sparse 50-dimensional signal. Here our focus is to enter the nonconvex recovery regime by comparing the convex ℓ_1-regularised with the nonconvex $\ell_{1/2}$-regularised problem. We discover that in either case reconstruction is feasible from already very few measurements. Increasing the number of optimising particles has almost no effect for the convex optimisation problem, in the nonconvex setting recovery benefits from more particles. Furthermore, the nonconvex problem demands a more moderate choice of the strength of the gradient.

5. Conclusions

In this paper, we investigate a variant of consensus-based optimisation (CBO) which incorporates memory effects and gradient information. By developing further and generalising the proof technique devised in [28, 29], we establish the global convergence of the underlying dynamics to the global minimiser x^* of the objective function $\mathcal{E}$ in mean-field law. To this end, we analyse the time-evolution of the Wasserstein
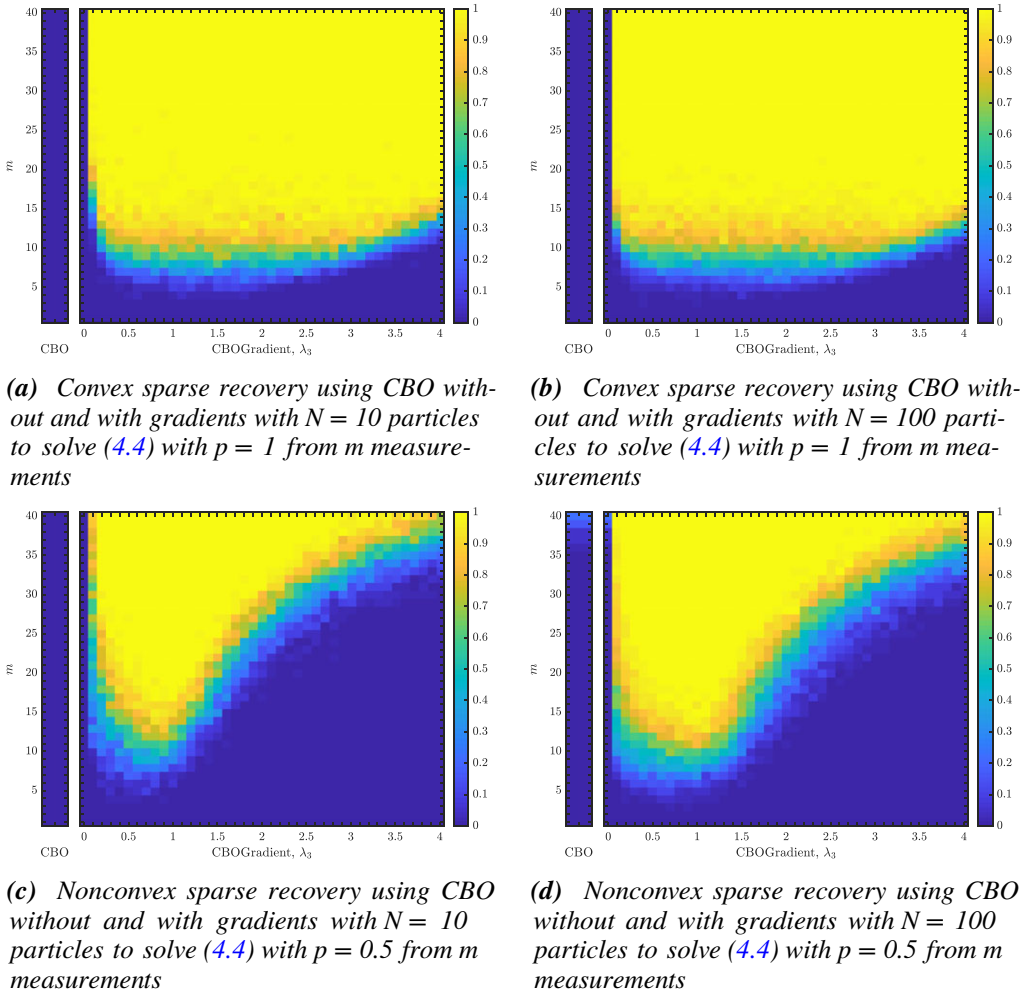
(a) *Convex sparse recovery using CBO without and with gradients with N = 10 particles to solve (4.4) with p = 1 from m measurements*

(b) *Convex sparse recovery using CBO without and with gradients with N = 100 particles to solve (4.4) with p = 1 from m measurements*

(c) *Nonconvex sparse recovery using CBO without and with gradients with N = 10 particles to solve (4.4) with p = 0.5 from m measurements*

(d) *Nonconvex sparse recovery using CBO without and with gradients with N = 100 particles to solve (4.4) with p = 0.5 from m measurements*

Figure 6. *Comparison between the success probabilities of CBO without (left separate columns) and with gradient information for different values of the parameter $\lambda_3 \in [0, 4]$ (right phase diagrams) with $N = 10$ ((**a**) and (**c**)) or $N = 100$ particles ((**b**) and (**d**)) when solving the convex or nonconvex ℓ_p-regularised least squares problem (4.4) with $p = 1$ and $\mu = $ ((**a**) and (**b**)) or $p = 0.5$ and $\mu = $ ((**c**) and (**d**)), respectively. On the vertical axis we depict the number of measurements m, from which we try to recover the 2-sparse and 50-dimensional sparse signal. As further parameters we choose the time horizon $T = 20$, discrete time step size $\Delta t = 0.01$, $\alpha = 100$, $\beta = \infty$, $\theta = 0$, $\kappa = 1/\Delta t$, $\lambda_1 = 1$, $\lambda_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_3 = 0$. We discover that in both the convex and nonconvex setting reconstruction is feasible from already very few measurements. While increasing the number of optimising particles has almost no effect for the convex optimisation problem, in the nonconvex setting recovery benefits from more particles. Note that the separate columns and the left most column of the phase diagrams coincide, and are only depicted in this way to highlight that we compare also CBO.*

distance between the law of the mean-field CBO dynamics and a Dirac delta at the minimiser and show its exponential decay in time. Our result holds under minimal assumptions about the initial measure ρ_0 and for a vast class of objective functions. The numerical benefit of such additional terms, specifically the employed memory effects and gradient information, is demonstrated at the example of a benchmark function in optimisation as well as at real-world applications such as compressed sensing and the training of neural networks for image classification.

By these means, we demonstrate the versatility, flexibility and customisability of the class of CBO methods, both with respect to potential application areas in practice and modifications to the underlying optimisation principles, while still being amenable to theoretical analysis.

Competing interests. None.

References

[1] Albi, G., Bongini, M., Cristiani, E. & Kalise, D. (2015) Invisible sparse control of self-organizing agents leaving unknown environments. *SIAM J. Appl. Math.* **76**(4), 1683–1710.

[2] Ambrosio, L., Gigli, N. & Savaré, G. (2008) *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics ETH Zürich, 2nd ed., Birkhäuser Verlag, Basel.

[3] Arnold, L. (1974) *Stochastic Differential Equations: Theory and Applications*, Wiley-Interscience [John Wiley & Sons], New York/London/Sydney. Translated from the German.

[4] Back, T., Fogel, D. B. & Michalewicz, Z. (1997) *Handbook of Evolutionary Computation*, Institute of Physics Publishing, Bristol; Oxford University Press, New York.

[5] Bae, H.-O., Ha, S.-Y., Kang, M., Lim, H., Min, C. & Yoo, J. (2022) A constrained consensus based optimization algorithm and its application to finance. *Appl. Math. Comput.* **416**, PaperNo.126726,10.

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

[7] Borghi, G., Herty, M. & Pareschi, L. (2022) A consensus-based algorithm for multi-objective optimization and its mean-field description. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, pp. 4131–4136.

[8] Borghi, G., Herty, M. & Pareschi, L. (2023) An adaptive consensus based method for multi-objective optimization with uniform Pareto front approximation. *Appl. Math. Optim.* **88**(2), 1–43.

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

[10] Bungert, L., Wacker, P. & Roith, T. (2022) Polarized consensus-based dynamics for optimization and sampling. *arXiv: 2211.05238*.

[11] Candès, E. J., Romberg, J. K. & Tao, T. (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**(8), 1207–1223.

[12] Carrillo, J. A., Choi, Y.-P., Totzeck, C. & Tse, O. (2018) An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.* **28**(6), 1037–1066.

[13] Carrillo, J. A., Hoffmann, F., Stuart, A. M. & Vaes, U. (2022) Consensus-based sampling. *Stud. Appl. Math.* **148**(3), 1069–1140.

[14] Carrillo, J. A., Jin, S., Li, L. & Zhu, Y. (2021) A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.* **27**, Paper No.S5,22.

[15] Carrillo, J. A., Totzeck, C. & Vaes, U. (2023) Consensus-based optimization and ensemble Kalman inversion for global optimization problems with constraints. In: *Modeling and Simulation for Collective Dynamics*, World Scientific, Singapore, pp. 195–230.

[16] Carrillo, J. A., Trillos, N. G., Li, S. & Zhu, Y. (2023). FedCBO: reaching group consensus in clustered federated learning through consensus-based optimization. *arXiv: 2305.02894*.

[17] Chen, J., Jin, S. & Lyu, L. (2020) A consensus-based global optimization method with adaptive momentum estimation. *arXiv: 2012.04827*.

[18] Cipriani, C., Huang, H. & Qiu, J. (2022) Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Appl. Math.* **54**(3), 3091–3121.

[19] Couzin, I. D., Krause, J., Franks, N. R. & Levin, S. A. (2005) Effective leadership and decision-making in animal groups on the move. *Nature* **433**(7025), 513–516.

[20] Cristiani, E., Piccoli, B. & Tosin, A. (2011) Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Model. Simul.* **9**(1), 155–182.

[21] Dembo, A. & Zeitouni, O. (1998) Large Deviations Techniques and Applications, *Applications of Mathematics (New York)*, Vol. 38, 2nd ed., Springer-Verlag, New York.

[22] Donoho, D. L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306.

[23] Dorigo, M. & Blum, C. (2005) Ant colony optimization theory: a survey. *Theor. Comput. Sci.* **344**(2-3), 243–278.

[24] Fogel, D. B. (2000) *Evolutionary Computation. Toward a New Philosophy of Machine Intelligence*, 2nd ed., IEEE Press, Piscataway, NJ.

[25] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2020) Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.* **30**(14), 2725–2751.

[26] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2021) Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.* **22**, PaperNo.237,55.

[27] Fornasier, M., Huang, H., Pareschi, L. & Sünnen, P. (2022) Anisotropic diffusion in consensus-based optimization on the sphere. *SIAM J. Optim.* **32**(3), 1984–2012.

[28] Fornasier, M., Klock, T. & Riedl, K. (2021). Consensus-based optimization methods converge globally. *arXiv: 2103.15130*.

[29] Fornasier, M., Klock, T. & Riedl, K. (2022) Convergence of anisotropic consensus-based optimization in mean-field law. In: Jiménez Laredo, J. L., Hidalgo, J. I. & Babaagba, K. O. (editors), *Applications of Evolutionary Computation*, Springer International Publishing, Cham, pp. 738–754.

[30] Foucart, S. & Rauhut, H. (2013) A mathematical introduction to compressive sensing. In: *Applied and Numerical Harmonic Analysis*, Birkhäuser/Springer, New York.

[31] Grassi, S., Huang, H., Pareschi, L. & Qiu, J. (2023). Mean-field particle swarm optimization. In: *Modeling and Simulation for Collective Dynamics* (pp. 127–193).

[32] Grassi, S. & Pareschi, L. (2021) From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. *Math. Models Methods Appl. Sci.* **31**(8), 1625–1657.

[33] Ha, S.-Y., Jin, S. & Kim, D. (2020) Convergence of a first-order consensus-based global optimization algorithm. *Math. Models Methods Appl. Sci.* **30**(12), 2417–2444.

[34] Ha, S.-Y., Jin, S. & Kim, D. (2021) Convergence and error estimates for time-discrete consensus-based optimization algorithms. *Numer. Math.* **147**(2), 255–282.

[35] Ha, S.-Y., Kang, M. & Kim, D. (2022) Emergent behaviors of high-dimensional Kuramoto models on Stiefel manifolds. *Automatica* **136**, PaperNo.110072.

[36] Hegselmann, R. & Krause, U. (2002) Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simul.* **5**(3).

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

[38] Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, University of Michigan Press, Ann Arbor, MI.

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

[40] Huang, H., Qiu, J. & Riedl, K. (2022) Consensus-based optimization for saddle point problems. *arXiv:2212.12334*.

[41] Huang, H., Qiu, J. & Riedl, K. (2023) On the global convergence of particle swarm optimization methods. *Appl. Math. Optim.* **88**(2), 30.

[42] Kalise, D., Sharma, A. & Tretyakov, M. V. (2023) Consensus-based optimization via jump-diffusion stochastic differential equations. *Math. Models Methods Appl. Sci.* **33**(02), 289–339.

[43] Kennedy, J. & Eberhart, R. (1995) Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*, Vol. **4**, IEEE, pp. 1942–1948.

[44] Kim, J., Kang, M., Kim, D., Ha, S.-Y. & Yang, I. (2020). A stochastic consensus method for nonconvex optimization on the Stiefel manifold. In: *2020 59th IEEE Conference on Decision and Control (CDC)*, IEEE, pp. 1050–1057.

[45] Klamroth, K., Stiglmayr, M. & Totzeck, C. (2022) Consensus-based optimization for multi-objective problems: a multi-swarm approach. *arXiv: 2211.15737*.

[46] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324.

[47] LeCun, Y., Cortes, C. & Burges, C. (2010) MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/

[48] Miller, P. D. (2006) *Applied Asymptotic Analysis*, Graduate Studies in Mathematics, Vol. **75**, American Mathematical Society, Providence, RI.

[49] Moulines, E. & Bach, F. (2011) Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: *Advances in Neural Information Processing Systems 24*.

[50] Øksendal, B. (2003) *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer-Verlag, Berlin.

[51] Parrish, J. K. & Edelstein-Keshet, L. (1999) Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science* **284**(5411), 99–101.

[52] Pinnau, R., Totzeck, C., Tse, O. & Martin, S. (2017) A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.* **27**(1), 183–204.

[53] Riedl, K., Klock, T., Geldhauser, C. & Fornasier, M. (2023) Gradient is all you need? *arXiv: 2306.09778*.

[54] Schillings, C., Totzeck, C. & Wacker, P. (2023) Ensemble-based gradient inference for particle methods in optimization and sampling. *SIAM/ASA J. Uncertain. Quantif.* **11**(3), 757–787.

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

[56] Vicsek, T. & Zafeiris, A. (2012) Collective motion. *Phys. Rep.* **517**(3-4), 71–140.