On the Global Convergence of Particle Swarm Optimization Methods

Hui Huang[1] · **Jinniao Qiu**[2] · **Konstantin Riedl**[3,4]

Abstract

In this paper we provide a rigorous convergence analysis for the renowned particle swarm optimization method by using tools from stochastic calculus and the analysis of partial differential equations. Based on a continuous-time formulation of the particle dynamics as a system of stochastic differential equations, we establish convergence to a global minimizer of a possibly nonconvex and nonsmooth objective function in two steps. First, we prove consensus formation of an associated mean-field dynamics by analyzing the time-evolution of the variance of the particle distribution, which acts as Lyapunov function of the dynamics. We then show that this consensus is close to a global minimizer by employing the asymptotic Laplace principle and a tractability condition on the energy landscape of the objective function. These results allow for the usage of memory mechanisms, and hold for a rich class of objectives provided certain conditions of well-preparation of the hyperparameters and the initial datum. In a second step, at least for the case without memory effects, we provide a quantitative result about the mean-field approximation of particle swarm optimization, which specifies the convergence of the interacting particle system to the associated mean-field limit. Combining these two results allows for global convergence guarantees of the numerical particle swarm optimization method with provable polynomial complexity. To demonstrate the applicability of the method we propose an efficient and parallelizable

---

✉ Konstantin Riedl
    konstantin.riedl@ma.tum.de

    Hui Huang
    hui.huang@uni-graz.at

    Jinniao Qiu
    jinniao.qiu@ucalgary.ca

[1] Institute of Mathematics and Scientific Computing, University of Graz, Graz, Austria

[2] Department of Mathematics and Statistics, University of Calgary, Calgary, Canada

[3] Department of Mathematics, School of Computation, Information and Technology, Technical University of Munich, Munich, Germany

[4] Munich Center for Machine Learning, Munich, Germany

implementation, which is tested in particular on a competitive and well-understood high-dimensional benchmark problem in machine learning.

Keywords Global derivative-free optimization · High-dimensional nonconvex optimization · Metaheuristics · Particle swarm optimization · Mean-field limit · Vlasov-Fokker-Planck equations

Mathematics Subject Classification 65C35 · 65K10 · 90C26 · 90C56 · 35Q90 · 35Q83

1 Introduction

In nature, collective behavior and self-organization allow complicated global patterns to emerge from simple interaction rules and random fluctuations. Inspired by the fascinating capabilities of swarm intelligence, large multi-agent systems are employed as a tool for solving challenging problems in applied mathematics. One classical task arising throughout science is concerned with the global optimization of a problem-dependent possibly nonconvex and nonsmooth objective function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$, i.e., the search for a global optimizer

$$x^* \in \arg \min_{x \in \mathbb{R}^d} \mathcal{E}(x). \tag{1.1}$$

A popular class of methods with a long history of achieving state-of-the-art performance on such problems are metaheuristics [14]. They orchestrate an interplay between local and global improvement procedures, consider memory mechanisms and selection strategies, and combine random and deterministic decisions, to create a process capable of escaping local optima and performing a robust search of the solution space in order to find a global optimizer. Initiated by seminal works on stochastic approximation [49] and random search [46], a big variety of such mechanisms has been introduced, analyzed and applied to numerous real-world problems. A non-exclusive list of representatives includes evolutionary programming [15], genetic algorithms [24], simulated annealing [1], and particle swarm optimization [31]. Despite their tremendous empirical success, it is very difficult to provide a theoretical convergence analysis to global minimizers, mostly due to their stochastic nature and the appearance of memory effects. Simulated annealing, however, is theoretically actually well-studied, see, e.g., the works [3, 37] as well as the recent survey [53] and references therein.

In this paper we study particle swarm optimization (PSO), which was initially introduced by Kennedy and Eberhart in the 90s [30, 31] and is now widely recognized as an efficient method for tackling complex optimization problems [35, 45]. Originally, PSO solves (1.1) by considering a group of finitely many particles, which explore the energy landscape of $\mathcal{E}$. Each agent experiences a force towards its own personal (historical) best position as well as towards the global best position communicated in the swarm. We refer to the ability of each particle remembering the best position it has been positioned at in the past as memory mechanisms. Although these interaction rules

are seemingly simple, a complete numerical analysis of PSO is still lacking; see, e.g., [41, 55, 57] and references therein. Recently, however, by introducing a continuous description of PSO based on a system of stochastic differential equations (SDEs), the authors of [22] have paved the way for a rigorous mathematical analysis using tools from stochastic calculus and the analysis of partial differential equations (PDEs).

In order to explore the domain and to form a global consensus about the minimizer x^* as time passes, the formulation of PSO proposed by the authors of [22] uses N particles, described by triplets $\left((X_t^i, Y_t^i, V_t^i)_{t \geq 0}\right)_{i=1 \ldots, N}$, with X_t^i and V_t^i denoting the position and velocity, and Y_t^i being a regularized version of the local (historical) best position, also referred to as memory, of the ith agent at time t. The particles, formally stochastic processes, are initialized independently according to some common distribution $f_0 \in \mathcal{P}(\mathbb{R}^{3d})$. In the most general form the PSO dynamics is given by the system of SDEs, expressed in Itô's form as

$$dX_t^i = V_t^i \, dt, \tag{1.2a}$$

$$dY_t^i = \kappa \left(X_t^i - Y_t^i\right) S^{\beta,\theta}\left(X_t^i, Y_t^i\right) dt, \tag{1.2b}$$

$$m \, dV_t^i = -\gamma V_t^i \, dt + \lambda_1\left(Y_t^i - X_t^i\right) dt + \lambda_2\left(y_\alpha(\widehat{\rho}_{Y,t}^N) - X_t^i\right) dt \tag{1.2c}$$
$$+ \sigma_1 D\left(Y_t^i - X_t^i\right) dB_t^{1,i} + \sigma_2 D\left(y_\alpha(\widehat{\rho}_{Y,t}^N) - X_t^i\right) dB_t^{2,i},$$

where $\alpha, \beta, \theta, \kappa, \gamma, m, \lambda_1, \lambda_2, \sigma_1, \sigma_2 \geq 0$ are user-specified parameters. The change of the velocity in (1.2c) is subject to five forces. The first term on the right-hand side models friction with a coefficient commonly chosen as $\gamma = 1 - m \geq 0$, where $m > 0$ denotes the inertia weight. The subsequent term can be regarded as the drift towards the local best position of the ith particle, which it has memorized in the state variable Y_t^i. A continuous-time approximation of its evolution is given by Y_t^i and described in Equation (1.2b). It involves the operator $S^{\beta,\theta}$, given by $S^{\beta,\theta}(x, y) = 1 + \theta + \tanh(\beta(\mathcal{E}(y) - \mathcal{E}(x)))$ for $0 \leq \theta \ll 1$ and $\beta \gg 1$, which converges to the Heaviside function as $\theta \to 0$ and $\beta \to \infty$. The concept behind Equation (1.2b) can then be seen when being discretized, see Remark 1. For an alternative implementation of the local best position we refer to [54].

Remark 1 A time-discretization of (1.2b) with $\kappa = 1/(2\Delta t)$, $\theta = 0$ and $\beta = \infty$ yields the update rule

$$Y_{(k+1)\Delta t}^i = \begin{cases} Y_{k\Delta t}^i, & \text{if } \mathcal{E}(X_{(k+1)\Delta t}^i) \geq \mathcal{E}(Y_{k\Delta t}^i), \\ X_{(k+1)\Delta t}^i, & \text{if } \mathcal{E}(X_{(k+1)\Delta t}^i) < \mathcal{E}(Y_{k\Delta t}^i), \end{cases} \tag{1.3}$$

meaning that the ith particle stores in $Y_{k\Delta t}^i$ the best position which it has seen up to the kth iteration. This explains the name local (historical) best position and restores the original definition from the work [31].

The last deterministic term imposes a drift towards the momentaneous consensus point $y_\alpha(\widehat{\rho}^N_{Y,t})$, given by

$$y_\alpha(\widehat{\rho}^N_{Y,t}) := \int_{\mathbb{R}^d} y \frac{\omega^{\mathcal{E}}_\alpha(y)}{\|\omega^{\mathcal{E}}_\alpha\|_{L_1(\widehat{\rho}^N_{Y,t})}} \, d\widehat{\rho}^N_{Y,t}(y), \quad \text{with} \quad \omega^{\mathcal{E}}_\alpha(y) := \exp(-\alpha\mathcal{E}(y)), \quad (1.4)$$

where $\widehat{\rho}^N_{Y,t}$ denotes the empirical measure $\widehat{\rho}^N_{Y,t} := \frac{1}{N}\sum_{i=1}^N \delta_{Y^i}$ of the particles' local best positions. The choice of the weight $\omega^{\mathcal{E}}_\alpha$ in (1.4) comes from the well-known Laplace principle [12, 38], a classical asymptotic argument for integrals stating that for any probability measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ it holds

$$\lim_{\alpha\to\infty} \left(-\frac{1}{\alpha} \log \left(\int_{\mathbb{R}^d} \omega^{\mathcal{E}}_\alpha(y) \, d\varrho(y) \right) \right) = \inf_{y\in\text{supp}(\varrho)} \mathcal{E}(y). \quad (1.5)$$

Based thereon, $y_\alpha(\widehat{\rho}^N_{Y,t})$ is expected to be a rough estimate for a global minimizer x^*, which improves as $\alpha \to \infty$ and as larger regions of the domain are explored. To feature the latter, the two remaining terms in (1.2c), each associated with a drift term, are diffusion terms injecting randomness into the dynamics through independent standard Brownian motions $((B^{1,i}_t)_{t\geq 0})_{i=1,\dots,N}$ and $((B^{2,i}_t)_{t\geq 0})_{i=1,\dots,N}$. The two commonly studied diffusion types for similar methods are isotropic [8, 18, 42] and anisotropic [9, 19] diffusion with

$$D(y-x) = \begin{cases} \|y-x\|_2 \, \text{Id}, & \text{for isotropic diffusion,} \\ \text{diag}\,(y-x), & \text{for anisotropic diffusion,} \end{cases} \quad (1.6)$$

where $\text{Id} \in \mathbb{R}^{d\times d}$ is the identity matrix and $\text{diag} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ the operator mapping a vector onto a diagonal matrix with the vector as its diagonal. Intuitively, the term's scaling encourages agents far from its own local best position or the globally computed consensus point to explore larger regions, whereas agents already close try to enhance their position only locally. As the coordinate-dependent scaling of anisotropic diffusion has been proven to be highly beneficial for high-dimensional problems [9, 17], in what follows, we limit our analysis to this case. An illustration of the formerly described PSO dynamics (1.2) is given in Fig. 1.

A theoretical convergence analysis of PSO is possible either on the microscopic level (1.2) or by analyzing the macroscopic behavior of the particle density through a mean-field limit, what usually admits more powerful analysis tools. In the large particle limit an individual particle is not influenced any more by individual particles but only by the average behavior of all particles. As shown in [21, Section 3.2], the empirical particle measure $\widehat{f}^N := \frac{1}{N}\sum_{i=1}^N \delta_{(X^i, Y^i, V^i)}$ converges in law to the deterministic agent distribution $f \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^{3d}))$, which weakly satisfies the nonlinear Vlasov-Fokker-Planck equation
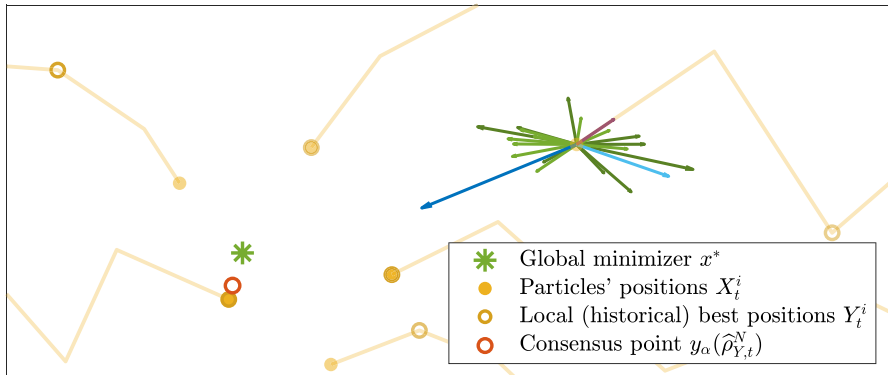
Fig. 1 An illustration of the PSO dynamics. Agents with positions $X^1, \ldots, X^N$ (yellow dots with their trajectories) explore the energy landscape of $\mathcal{E}$ in search of the global minimizer x^* (green star). The dynamics of each particle is governed by five terms. A local drift term (light blue arrow) imposes a force towards its local best position Y_t^i (indicated by a circle). A global drift term (dark blue arrow) drags the agent towards a momentaneous consensus point $y_\alpha(\widehat{\rho}_{Y,t}^N)$ (orange circle) computed as a weighted (visualized through color opacity) average of the particles' local best positions. Friction (purple arrow) counteracts inertia. The two remaining terms are diffusion terms (light and dark green arrows) associated with a respective drift term

$$\partial_t f_t + v \cdot \nabla_x f_t + \nabla_y \cdot \left(\kappa (x - y) S^{\beta,\theta}(x, y) f_t \right)$$

$$= \nabla_v \cdot \left(\frac{\gamma}{m} v f_t + \frac{\lambda_1}{m} (x - y) f_t + \frac{\lambda_2}{m} \left(x - y_\alpha(\rho_{Y,t}) \right) f_t \right.$$

$$\left. + \left(\frac{\sigma_1^2}{2m^2} \left(D(x - y) \right)^2 + \frac{\sigma_2^2}{2m^2} \left(D(x - y_\alpha(\rho_{Y,t})) \right)^2 \right) \nabla_v f_t \right) \quad (1.7)$$

with initial datum f_0. The mean-field limit results [6, 25–27, 52] ensure that the particle system (1.2) is well-approximated by the following self-consistent mean-field McKean process

$$d\overline{X}_t = \overline{V}_t \, dt, \tag{1.8a}$$

$$d\overline{Y}_t = \kappa \left(\overline{X}_t - \overline{Y}_t \right) S^{\beta,\theta} \left(\overline{X}_t, \overline{Y}_t \right) dt, \tag{1.8b}$$

$$m \, d\overline{V}_t = -\gamma \overline{V}_t \, dt + \lambda_1 \left(\overline{Y}_t - \overline{X}_t \right) dt + \lambda_2 \left(y_\alpha(\rho_{Y,t}) - \overline{X}_t \right) dt \tag{1.8c}$$
$$+ \sigma_1 D \left(\overline{Y}_t - \overline{X}_t \right) dB_t^1 + \sigma_2 D \left(y_\alpha(\rho_{Y,t}) - \overline{X}_t \right) dB_t^2,$$

with initial datum $(\overline{X}_0, \overline{Y}_0, \overline{V}_0) \sim f_0$ and the marginal law $\rho_{Y,t}$ of $\overline{Y}_t$ given by

$$\rho_Y(t, \cdot) = \iint_{\mathbb{R}^d \times \mathbb{R}^d} df_t(x, \cdot, v).$$

Here, f_t denotes the distribution of $(\overline{X}_t, \overline{Y}_t, \overline{V}_t)$. This makes (1.7) and (1.8) nonlinear.

1.1 Contribution

In view of the versatility, efficiency, and wide applicability of PSO combined with its long historical tradition, a mathematical analysis of the finite particle system (1.2) is of considerable interest.

In this work we advance the theoretical understanding of the method and contribute to the completion of a full numerical analysis of PSO by proving rigorously the convergence of PSO with memory effects to global minimizers using mean-field techniques. More precisely, under mild regularity assumptions on the objective $\mathcal{E}$ and a well-preparation condition about the initialization f_0, we analyze the behavior of the particle distribution f solving the mean-field dynamics (1.8). At first, it is shown that concentration is achieved at some $\widetilde{x}$ in the sense that the marginal law w.r.t. the local best position, $\rho_{Y,t}$, converges narrowly to a Dirac delta $\delta_{\widetilde{x}}$ as $t \to \infty$. Consecutively, we argue that, for an appropriate choice of the parameters, in particular $\alpha \gg 1$, which may depend on the dimension d, $\mathcal{E}(\widetilde{x})$ can be made arbitrarily close to the minimal value $\underline{\mathcal{E}} := \inf_{x \in \mathbb{R}^d} \mathcal{E}(x)$. A suitable tractability condition on the objective $\mathcal{E}$ eventually ensures that $\widetilde{x}$ is close to a global minimizer. Similar mean-field convergence results are obtained for the case without memory effects. In this setting we are moreover able to establish the convergence of the interacting N-particle dynamics to its mean-field limit with a dimension-independent rate, which allows to obtain a so far unique holistic and quantitative convergence statement of PSO. As the mean-field approximation result does not suffer from the curse of dimensionality, we in particular prove that the numerical PSO method has polynomial complexity. With these new results we solve the theoretical open problem about the convergence of PSO posed in [22].

Furthermore, we propose an efficient and parallelizable implementation, which is particularly suited for machine learning problems by integrating modern machine learning techniques such as random mini-batch ideas as well as traditional metaheuristic-inspired techniques from genetic programming and simulated annealing.

1.2 Prior Arts

The convergence of PSO algorithms has been investigated by many scholars since its introduction, which has lead to several variations allowing to establish desirable properties such as consensus formation or convergence to optimal solutions. While the matter of consensus is well-studied, see, e.g., [11, 40] or more recently [56], where the authors employ stochastic approximation methods [32], only few general theoretical statements regarding the properties of the found consensus are available. Both the existence of a large number of variations of the algorithm and the lack of a rigoros global convergence analysis are attributed amongst other things, such as the stochasticity and the usage of memory mechanisms, to the phenomenon of premature convergence of basic PSO [31], which was observed in [4, 5] and remedied by proposing a modified version, called guaranteed convergence PSO. Nevertheless, this adaptation only allows to prove the convergence to local optima. In order to obtain therefrom a stochastic global search algorithm, the authors suggest to add

purely stochastic particles to the swarm, which trivially makes the method capable of detecting a global optimizer, but entails a computational time which coincides with the time required to examine every location in the search space. Other works consider certain notions of weak convergence [7] or provide probabilistic guarantees of finding locally optimal solutions, meaning that eventually all particles are located almost surely at a local optimum of the objective function [51]. In [44], similarly to our work, the expected behavior of the particles is investigated.

All of the formerly mentioned results though are obtained through the analysis of the particles' trajectories generated by a time-discretized algorithm as in [21, Equation (6.3)]. The present paper takes a different point of view by studying the continuous-time description of the PSO model (1.2) through the lens of the mean-field approximation (1.7). Analyzing the macroscopic behavior of a system through a mean-field limit instead of investigating the microscopic particle dynamics has its origins in statistical mechanics [29], where interactions between particles are approximated by an averaged influence. By eliminating the correlation between the particles, a many-body problem can be reduced to a one-body problem, which is usually much easier to solve while still giving an understanding of the mechanisms at play by describing the average behavior of the particles. These ideas, for instance, are also used to study the collective behavior of animals when forming large-scale patterns through self-organization by analyzing an associated kinetic PDE [6]. In very recent works, this perspective of analysis has also been taken to demystify the training process of neural networks, see, e.g., [13, 36], where a mean-field approximation is utilized to formulate risk minimization by stochastic gradient descent (SGD) in terms of a gradient-flow PDE, which allows for a rigorous mathematical analysis.

The analysis technique we use follows the line of work of self-organization. It is inspired by [8, 9], where a variance-based analysis approach has been developed for consensus-based optimization (CBO), which follows the guiding principles of meta-heuristics and in particular resembles PSO but is of much simpler nature and therefore easier to analyze. In comparison to Equation (1.2), CBO methods are described by a system of first-order SDEs [8, Equation (1.1)] and do not contain memory mechanisms, which are responsible for both a significantly more challenging mathematical modeling and convergence analysis.

1.3 Organization

Sections 2 and 3 are dedicated to the analysis of PSO without and with memory mechanisms, respectively. After providing details about the well-posedness of the mean-field dynamics, we present and discuss the main result about the convergence of the mean-field dynamics to a global minimizer of the objective function. In Sect. 4 we then state a quantitative result about the mean-field approximation for PSO without memory effects, which enables us to obtain a holistic convergence statement of the numerical PSO method. Eventually, a computationally efficient implementation of PSO is proposed in Sect. 5, before Sect. 6 concludes the paper. For the sake of reproducible research, in the GitHub repository https://github.com/KonstantinRiedl/PSOAnalysis we provide the Matlab code implementing the PSO algorithm analyzed in this work.

2 Mean-Field Analysis of PSO Without Memory Effects

Before providing a theoretical analysis of the mean-field PSO dynamics (1.7) and (1.8), in this section we investigate a reduced version, which does not involve memory mechanisms. Its multi-particle formulation was proposed in [22, Section 3.1] and reads

$$dX_t^i = V_t^i \, dt, \tag{2.1a}$$

$$m \, dV_t^i = -\gamma V_t^i \, dt + \lambda \left(x_\alpha(\widehat{\rho}_{X,t}^N) - X_t^i \right) dt + \sigma D \left(x_\alpha(\widehat{\rho}_{X,t}^N) - X_t^i \right) dB_t^i. \tag{2.1b}$$

Compared to the full model, each particle is characterized only by its position X^i and velocity V^i. The forces acting on a particle, i.e., influencing its velocity in Equation (2.1b), are friction, acceleration through the consensus drift and diffusion as in (1.6) with independent standard Brownian motions $\left((B_t^i)_{t \geq 0}\right)_{i=1,\dots,N}$. The consensus point $x_\alpha(\widehat{\rho}_{X,t}^N)$ is directly computed from the current positions of the particles according to

$$x_\alpha(\widehat{\rho}_{X,t}^N) := \int_{\mathbb{R}^d} x \, \frac{\omega_\alpha^{\mathcal{E}}(x)}{\left\| \omega_\alpha^{\mathcal{E}} \right\|_{L_1(\widehat{\rho}_{X,t}^N)}} \, d\widehat{\rho}_{X,t}^N(x), \tag{2.2}$$

where $\widehat{\rho}_{X,t}^N$ denotes the empirical measure $\widehat{\rho}_{X,t}^N := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$ of the particles' positions. Independent and identically distributed initial data $\left((X_0^i, V_0^i) \sim f_0\right)_{i=1,\dots,N}$ with $f_0 \in \mathcal{P}(\mathbb{R}^{2d})$ complement (2.1).

Similar to the particle system (1.2), as $N \to \infty$, the mean-field dynamics of (2.1) is described by the nonlinear self-consistent McKean process

$$d\overline{X}_t = \overline{V}_t \, dt, \tag{2.3a}$$

$$m \, d\overline{V}_t = -\gamma \overline{V}_t \, dt + \lambda \left(x_\alpha(\rho_{X,t}) - \overline{X}_t \right) dt + \sigma D \left(x_\alpha(\rho_{X,t}) - \overline{X}_t \right) dB_t, \tag{2.3b}$$

with initial datum $(\overline{X}_0, \overline{V}_0) \sim f_0$ and the marginal law $\rho_{X,t}$ of $\overline{X}_t$ given by $\rho_X(t, \cdot) = \int_{\mathbb{R}^d} df(t, \cdot, v)$. A direct application of the Itô-Doeblin formula shows that the law $f \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^{2d}))$ is a weak solution to the nonlinear Vlasov-Fokker-Planck equation

$$\partial_t f_t + v \cdot \nabla_x f_t$$
$$= \nabla_v \cdot \left(\frac{\gamma}{m} v f_t + \frac{\lambda}{m} \left(x - x_\alpha(\rho_{X,t}) \right) f_t + \frac{\sigma^2}{2m^2} \left(D(x - x_\alpha(\rho_{X,t})) \right)^2 \nabla_v f_t \right) \tag{2.4}$$

with initial datum f_0.

Remark 2 A separate theoretical analysis of the dynamics (2.1) is necessary as it cannot be derived from (1.2) in a way that also the proof technique can be adopted in a straightforward manner. This can be seen from subtle differences in the proofs of Theorems 2 and 4; see in particular Lemma 3.

It is also worth noting that Equation (2.1) bears a certain resemblance to CBO [8, 9, 18, 19, 42], whereas (1.8) resembles [48]. Indeed, as made rigorous in [10], CBO methods can be derived from PSO in the small inertia limit $m \to 0$, or equivalently $\gamma \to 1$. Nevertheless, analyzing the convergence of CBO directly permits sharper bounds when compared to utilizing the results obtained in our work together with [10, Theorem 2.4].

Before turning towards the well-posedness of the mean-field dynamics (2.3) and presenting the main result of this section about the convergence to the global minimizer x^*, let us introduce the class of objective function $\mathcal{E}$ considered in the theoretical part of this work. We remark that the assumptions made in what follows coincide with the ones of [8, 9] as well as several subsequent works in this direction.

Assumption 1 Throughout the paper we are interested in objective functions $\mathcal{E}$: $\mathbb{R}^d \to \mathbb{R}$, for which

A1 there exists $x^* \in \mathbb{R}^d$ such that $\mathcal{E}(x^*) = \inf_{x \in \mathbb{R}^d} \mathcal{E}(x) =: \underline{\mathcal{E}}$,
A2 there exists some constant $L_{\mathcal{E}} > 0$ such that

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

A3 either $\overline{\mathcal{E}} := \sup_{x \in \mathbb{R}^d} \mathcal{E}(x) < \infty$ or there exist constants $c_{\mathcal{E}}, R > 0$ such that

$$\mathcal{E}(x) - \underline{\mathcal{E}} \geq c_{\mathcal{E}} |x|^2, \quad \text{for all } x \in \mathbb{R}^d \text{ with } |x| \geq R,$$

A4 $\mathcal{E} \in \mathcal{C}^2(\mathbb{R}^d)$ with $\left\| \nabla^2 \mathcal{E} \right\|_\infty \leq C_{\mathcal{E}}$ for some constant $C_{\mathcal{E}} > 0$,
A5 there exist $\eta > 0$ and $\nu \in (0, \infty)$ such that for any $x \in \mathbb{R}^d$ there exists a global minimizer x^* of $\mathcal{E}$ (which may depend on x) such that

$$\left| x - x^* \right| \leq (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu / \eta.$$

Assumption A1 just states that the objective function $\mathcal{E}$ attains its infimum $\underline{\mathcal{E}}$ at some $x^* \in \mathbb{R}^d$, which may not necessarily be unique. Assumption A2 describes the local Lipschitz-continuity of $\mathcal{E}$, entailing in particular that the objective has at most quadratic growth at infinity. Assumption A3, on the other hand, requires $\mathcal{E}$ to be either bounded or of at least quadratic growth in the farfield. Together, A2 and A3 allow to obtain the well-posedness of the PSO model. Assumption A4 is a regularity assumption about $\mathcal{E}$, which is required only for the theoretical analysis. The quadratic growth nature of Assumptions A2–A4 in the farfield may bear a certain resemblance to log-Sobolev inequalities [50], which are pivotal in the convergence analysis of simulated annealing, see [53] for further details. Unlike simulated annealing however, the PSO method is a zero-order method where we do not need the gradient information of the objective function in the numerical application. Assumption A5 should be interpreted as a tractability condition of the landscape of $\mathcal{E}$, which ensures that achieving an objective value of approximately $\underline{\mathcal{E}}$ guarantees closeness to a global minimizer x^* and thus eliminates cases of almost-optimal valleys in the energy landscape far away from

any globally minimizing argument. Such assumption is therefore also referred to as an inverse continuity property.

It shall be emphasized that objectives with multiple global minima of identical quality are not excluded.

2.1 Well-Posedness of PSO without Memory Effects

Let us recite a well-posedness result about the mean-field PSO dynamics (2.3) and the associated Vlasov-Fokker-Planck equation (2.4). Its proof is analogous to the one provided for Theorem 3 for the full dynamics (1.8) based on the Leray-Schauder fixed point theorem.

Theorem 1 *Let $\mathcal{E}$ satisfy Assumptions A1–A3. Moreover, let $m, \gamma, \lambda, \sigma, \alpha, T > 0$. If $(\overline{X}_0, \overline{V}_0)$ is distributed according to $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$, then the nonlinear SDE (2.3) admits a unique strong solution up to time T with the paths of process $(\overline{X}, \overline{V})$ valued in $\mathcal{C}([0, T], \mathbb{R}^d) \times \mathcal{C}([0, T], \mathbb{R}^d)$. The associated law f has regularity $\mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ and is a weak solution to the Vlasov-Fokker-Planck equation (2.4). In particular,*

$$\sup_{t \in [0,T]} \mathbb{E}[|\overline{X}_t|^4 + |\overline{V}_t|^4] \leq \left(1 + 2\mathbb{E}[|\overline{X}_0|^4 + |\overline{V}_0|^4] \right) e^{CT} \qquad (2.5)$$

for some constant $C > 0$ depending only on $m, \gamma, \lambda, \sigma, \alpha, c_{\mathcal{E}}, R$, and $L_{\mathcal{E}}$.

2.2 Convergence of PSO without Memory Effects to a Global Minimizer

A successful application of the PSO dynamics underlies the premise that the particles form consensus about a certain position $\widetilde{x}$. In particular, in the mean-field limit one expects that the distribution of a particle's position $\rho_{X,t}$ converges to a Dirac delta $\delta_{\widetilde{x}}$. This entails that the variance in the position $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]$ and the second-order moment of the velocity $\mathbb{E}[|\overline{V}_t|^2]$ of the averaged particle vanish. As we show in what follows, both functionals indeed decay exponentially fast in time. Motivated by these expectations we define the functional

$$\mathcal{H}(t) := \left(\frac{\gamma}{2m} \right)^2 |\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2 + \frac{\gamma}{2m} \left\langle \overline{X}_t - \mathbb{E}[\overline{X}_t], \overline{V}_t \right\rangle, \qquad (2.6)$$

which we analyze in the remainder of this section. Its last term is required from a technical perspective. However, by proving the decay of $\mathbb{E}[\mathcal{H}(t)]$, which acts as Lyapunov function of the dynamics, one immediately obtains the same for $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2]$ as a consequence of the equivalence established in Lemma 1, which follows from Young's inequality.

Lemma 1 *The functional $\mathcal{H}(t)$ is equivalent to $|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2$ in the sense that*

$$
\frac{1}{2}\left(\frac{\gamma}{2m}\right)^2 |\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + \frac{1}{2}|\overline{V}_t|^2
$$
$$
\leq \mathcal{H}(t) \leq \frac{3}{2}\left(\left(\frac{\gamma}{2m}\right)^2 + 1\right)\left(|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2\right). \tag{2.7}
$$

We now derive an evolution inequality of the Lyapunov function $\mathbb{E}[\mathcal{H}(t)]$.

Lemma 2 *Let $\mathcal{E}$ satisfy Assumptions A1–A3 and let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be a solution to the nonlinear SDE (2.3). Then $\mathbb{E}[\mathcal{H}(t)]$ with $\mathcal{H}$ as defined in (2.6) satisfies*

$$
\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \leq -\frac{\gamma}{m}\mathbb{E}[|\overline{V}_t|^2]
$$
$$
-\left(\frac{\lambda\gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2}\right)\frac{2e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]}\right)\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]. \tag{2.8}
$$

Proof Let us write $\delta\overline{X}_t := \overline{X}_t - \mathbb{E}[\overline{X}_t]$ for short and note that the integration by parts formula gives

$$
\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] = 2\mathbb{E}[\langle\delta\overline{X}_t, \overline{V}_t\rangle]. \tag{2.9}
$$

Observe that, in what follows, the appearing stochastic integrals have vanishing expectations as a consequence of the regularity $f \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ obtained in Theorem 1. This is due to [39, Theorem 3.2.1(iii), Definition 3.1.4(iii)], which state that a stochastic integral vanishes if the associated second moment is integrable. Notice that the latter condition is sufficient for the stochastic integral to be a martingale. Applying the Itô-Doeblin formula and Young's inequality yields

$$
\frac{d}{dt}\mathbb{E}[|\overline{V}_t|^2] = -\frac{2\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] + \frac{2\lambda}{m}\mathbb{E}[\langle\overline{V}_t, x_\alpha(\rho_{X,t}) - \overline{X}_t\rangle] + \frac{\sigma^2}{m^2}\mathbb{E}[|x_\alpha(\rho_{X,t}) - \overline{X}_t|^2]
$$
$$
\leq -\left(\frac{2\gamma}{m} - \frac{\lambda}{\varepsilon m}\right)\mathbb{E}[|\overline{V}_t|^2] + \left(\frac{\varepsilon\lambda}{m} + \frac{\sigma^2}{m^2}\right)\mathbb{E}[|x_\alpha(\rho_{X,t}) - \overline{X}_t|^2], \quad \forall \varepsilon > 0. \tag{2.10}
$$

Again by employing the Itô-Doeblin formula we obtain

$$
\frac{d}{dt}\mathbb{E}[\langle\delta\overline{X}_t, \overline{V}_t\rangle] = \mathbb{E}[|\overline{V}_t|^2] - (\mathbb{E}[\overline{V}_t])^2 - \frac{\gamma}{m}\mathbb{E}[\langle\delta\overline{X}_t, \overline{V}_t\rangle] + \frac{\lambda}{m}\mathbb{E}[\langle\delta\overline{X}_t, x_\alpha(\rho_{X,t}) - \overline{X}_t\rangle]
$$
$$
\leq \mathbb{E}[|\overline{V}_t|^2] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] + \frac{\lambda}{m}\mathbb{E}[\langle\delta\overline{X}_t, x_\alpha(\rho_{X,t}) - \mathbb{E}[\overline{X}_t]\rangle] - \frac{\lambda}{m}\mathbb{E}[|\delta\overline{X}_t|^2] \tag{2.11}
$$
$$
= \mathbb{E}[|\overline{V}_t|^2] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] - \frac{\lambda}{m}\mathbb{E}[|\delta\overline{X}_t|^2],
$$

where we used the identity (2.9) and the fact that $\mathbb{E}[\langle \delta \overline{X}_t, x_\alpha(\rho_{X,t}) - \mathbb{E}[\overline{X}_t]\rangle] = 0$ in the last two steps. We now rearrange inequality (2.11) to get

$$\frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] + \frac{d}{dt}\mathbb{E}[\langle \delta\overline{X}_t, \overline{V}_t\rangle] \leq \mathbb{E}[|\overline{V}_t|^2] - \frac{\lambda}{m}\mathbb{E}[|\delta\overline{X}_t|^2],$$

which, in combination with (2.10), allows to show

$$\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \leq -\left(\frac{3\gamma}{2m} - \frac{\lambda}{\varepsilon m}\right)\mathbb{E}[|\overline{V}_t|^2] - \frac{\lambda\gamma}{2m^2}\mathbb{E}[|\delta\overline{X}_t|^2]$$

$$+ \left(\frac{\varepsilon\lambda}{m} + \frac{\sigma^2}{m^2}\right)\mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2]. \tag{2.12}$$

In order to upper bound $\mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2]$, an application of Jensen's inequality yields

$$\mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2] \leq \frac{\iint_{\mathbb{R}^{2d}}|x - x'|^2\,\omega_\alpha^{\mathcal{E}}(x')\,d\rho_{X,t}(x')d\rho_{X,t}(x)}{\int_{\mathbb{R}^d}\omega_\alpha^{\mathcal{E}}(x')\,d\rho_{X,t}(x')}$$

$$\leq 2e^{-\alpha\underline{\mathcal{E}}}\frac{\mathbb{E}[|\delta\overline{X}_t|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]}. \tag{2.13}$$

By choosing $\varepsilon = (2\lambda)/\gamma$ in (2.12) and utilizing the estimate (2.13), we obtain (2.8) as desired. $\qquad\square$

Remark 3 To obtain exponential decay of $\mathbb{E}[\mathcal{H}(t)]$ it is necessary to ensure the negativity of the prefactor of $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]$ in Inequality (2.8) by choosing the parameters of the PSO method in a suitable manner. This may be achieved by choosing for any fixed time t, given α and arbitrary $\sigma, \gamma > 0$,

$$\lambda > 4D_t^X\sigma^2/\gamma \quad \text{and subsequently} \quad m < \gamma^2/(8D_t^X\lambda), \tag{2.14}$$

where we abbreviate $D_t^X = 2e^{-\alpha\underline{\mathcal{E}}}/\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]$.

In order to be able to choose the parameters in Remark 3 once at the beginning of the algorithm instead of at every time step t, we need to be able to control the time-evolution of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]$. We therefore study its time-derivative in the following lemma.

Lemma 3 *Let $\mathcal{E}$ satisfy Assumptions A1–A4 and let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be the solution to the nonlinear SDE (2.3). Then it holds that*

$$\frac{d^2}{dt^2}\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2 \geq -\frac{\gamma}{m}\frac{d}{dt}\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2$$

$$-4\alpha e^{-2\alpha\underline{\mathcal{E}}}C_{\mathcal{E}}\left(1 + 2\frac{\lambda}{m}\left(\frac{2m}{\gamma}\right)^2\right)\mathbb{E}[\mathcal{H}(t)]. \tag{2.15}$$

Proof We first note that

$$
\begin{aligned}
\frac{1}{2}\frac{d^2}{dt^2} & \left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2 \\
&= \frac{d}{dt}\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\frac{d}{dt}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right) \\
&= \left(\frac{d}{dt}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2 \\
&\quad + \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\frac{d^2}{dt^2}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))] \\
&\geq \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\frac{d^2}{dt^2}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))],
\end{aligned}
\tag{2.16}
$$

leaving the second time-derivative of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]$ to be lower bounded. To do so, we start with its first derivative. Applying the Itô-Doeblin formula twice and noting that stochastic integrals have vanishing expectations as a consequence of [39, Theorem 3.2.1(iii), Definition 3.1.4(iii)] combined with the regularity $f \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{2d}))$ obtained in Theorem 1, we have

$$
\begin{aligned}
\frac{d}{dt}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))] &= -\alpha\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))\langle\nabla\mathcal{E}(\overline{X}_t), \overline{V}_t\rangle] \\
&= -\alpha\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))\langle\nabla\mathcal{E}(\overline{X}_0), \overline{V}_0\rangle] \\
&\quad - \alpha\mathbb{E}\left[\int_0^t d\left\langle\exp(-\alpha\mathcal{E}(\overline{X}_s))\nabla\mathcal{E}(\overline{X}_s), \overline{V}_s\right\rangle\right] \\
&= -\alpha\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))\langle\nabla\mathcal{E}(\overline{X}_0), \overline{V}_0\rangle] \\
&\quad - \alpha\mathbb{E}\left[\int_0^t \left\langle\exp(-\alpha\mathcal{E}(\overline{X}_s))\overline{V}_s, \nabla^2\mathcal{E}(\overline{X}_s)\overline{V}_s\right\rangle ds\right] \\
&\quad + \alpha^2\mathbb{E}\left[\int_0^t \exp(-\alpha\mathcal{E}(\overline{X}_s))\left|\langle\nabla\mathcal{E}(\overline{X}_s), \overline{V}_s\rangle\right|^2 ds\right] \\
&\quad - \alpha\mathbb{E}\left[\int_0^t \exp(-\alpha\mathcal{E}(\overline{X}_s))\left\langle\nabla\mathcal{E}(\overline{X}_s), -\frac{\gamma}{m}\overline{V}_s\right\rangle ds\right] \\
&\quad - \alpha\mathbb{E}\left[\int_0^t \exp(-\alpha\mathcal{E}(\overline{X}_s))\left\langle\nabla\mathcal{E}(\overline{X}_s), \frac{\lambda}{m}(x_\alpha(\rho_{X,s}) - \overline{X}_s)\right\rangle ds\right].
\end{aligned}
\tag{2.17}
$$

Differentiating both sides of (2.17) with respect to the time t yields

$$
\begin{aligned}
\frac{d^2}{dt^2}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))] &= -\alpha\mathbb{E}[\langle\exp(-\alpha\mathcal{E}(\overline{X}_t))\overline{V}_t, \nabla^2\mathcal{E}(\overline{X}_t)\overline{V}_t\rangle] \\
&\quad + \alpha^2\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))|\langle\nabla\mathcal{E}(\overline{X}_t), \overline{V}_t\rangle|^2] \\
&\quad + \frac{\alpha\gamma}{m}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))\langle\nabla\mathcal{E}(\overline{X}_t), \overline{V}_t\rangle]
\end{aligned}
$$

$$- \frac{\alpha\lambda}{m}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))\langle\nabla\mathcal{E}(\overline{X}_t), x_\alpha(\rho_{X,t}) - \overline{X}_t\rangle]$$

$$\geq -\frac{\gamma}{m}\frac{d}{dt}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]$$

$$- \alpha\underbrace{\mathbb{E}[\langle\exp(-\alpha\mathcal{E}(\overline{X}_t))\overline{V}_t, \nabla^2\mathcal{E}(\overline{X}_t)\overline{V}_t\rangle]}_{T_1}$$

$$- \frac{\alpha\lambda}{m}\underbrace{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))\langle\nabla\mathcal{E}(\overline{X}_t), x_\alpha(\rho_{X,t}) - \overline{X}_t\rangle]}_{T_2},$$

$$(2.18)$$

where we employed the first line of (2.17) in the last step. It remains to upper bound the terms T_1 and T_2. Making use of Assumptions A1 and A4, we immediately obtain

$$T_1 \leq \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))\|\nabla^2\mathcal{E}\|_\infty|\overline{V}_t|^2] \leq e^{-\alpha\underline{\mathcal{E}}}C_\mathcal{E}\mathbb{E}[|\overline{V}_t|^2]. \quad (2.19)$$

For T_2, again under Assumptions A1 and A4, we first note that

$$T_2 = -\mathbb{E}\big[\exp(-\alpha\mathcal{E}(\overline{X}_t))\big\langle\nabla\mathcal{E}(\overline{X}_t) - \nabla\mathcal{E}(x_\alpha(\rho_{X,t})), \overline{X}_t - x_\alpha(\rho_{X,t})\big\rangle\big]$$
$$\leq e^{-\alpha\underline{\mathcal{E}}}C_\mathcal{E}\mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2], \quad (2.20)$$

where the equality is a consequence of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))\langle\nabla\mathcal{E}(x_\alpha(\rho_{X,t})), \overline{X}_t - x_\alpha(\rho_{X,t})\rangle] = 0$, which follows from the definition of $x_\alpha(\rho_{X,t})$. Bounding $\mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2]$ as in (2.13) we can further bound (2.20) as

$$T_2 \leq e^{-\alpha\underline{\mathcal{E}}}C_\mathcal{E}\mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2] \leq 2e^{-2\alpha\underline{\mathcal{E}}}C_\mathcal{E}\frac{\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]}. \quad (2.21)$$

Collecting the estimates (2.19) and (2.21) within (2.18) and inserting the result into (2.16) give

$$\frac{1}{2}\frac{d^2}{dt^2}\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2 \geq -\frac{\gamma}{m}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\frac{d}{dt}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]$$

$$- \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\alpha C_\mathcal{E}e^{-\alpha\underline{\mathcal{E}}}\mathbb{E}[|\overline{V}_t|^2]$$

$$- \frac{2\alpha\lambda}{m}e^{-2\alpha\underline{\mathcal{E}}}C_\mathcal{E}\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]$$

$$\geq -\frac{\gamma}{2m}\frac{d}{dt}\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2$$

$$- \alpha e^{-2\alpha\underline{\mathcal{E}}}C_\mathcal{E}\left(\mathbb{E}[|\overline{V}_t|^2] + \frac{2\lambda}{m}\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]\right),$$

which yields the statement after employing the lower bound of (2.7) as in Lemma 1.

$\square$

We are now ready to state and prove the main result about the convergence of the mean-field PSO dynamics (2.3) without memory mechanisms to the global minimizer x^*.

Theorem 2 *Let $\mathcal{E}$ satisfy Assumptions A1–A4 and let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be a solution to the nonlinear SDE (2.3). Moreover, let us assume the well-preparation of the initial datum $\overline{X}_0$ and $\overline{V}_0$ in the sense that*

P1 $\mu > 0$ with

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

P2 it holds

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

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

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

Then $\mathbb{E}[\mathcal{H}(t)]$ with $\mathcal{H}$ as defined in Equation (2.6) converges exponentially fast with rate χ to 0 as $t \to \infty$. Moreover, there exists some $\widetilde{x}$, which may depend on α and f_0, such that $\mathbb{E}[\overline{X}_t] \to \widetilde{x}$ and $x_\alpha(\rho_{X,t}) \to \widetilde{x}$ exponentially fast with rate $\chi/2$ as $t \to \infty$. Eventually, for any given accuracy $\varepsilon > 0$, there exists $\alpha_0 > 0$, which may depend on the dimension d, such that for all $\alpha > \alpha_0$, $\widetilde{x}$ satisfies

$$\mathcal{E}(\widetilde{x}) - \underline{\mathcal{E}} \leq \varepsilon.$$

If $\mathcal{E}$ additionally satisfies Assumption A5, we have $|\widetilde{x} - x^*| \leq \varepsilon^\nu/\eta$.

Remark 4 As suggested in Remark 3, Theorem 2 traces back the evolution of $\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]$ to its initial state by employing Lemma 3. This allows to fixate all parameters of PSO at initialization time. By replacing D_t^X with $2D_0^X$ in (2.14), the well-preparation of the parameters as in Condition P1 can be ensured.

Condition P2 requires the well-preparation of the initialization in the sense that the initial datum f_0 is both well-concentrated and to a certain extent not too far from an optimal value. While this might have a locality flavor, the condition is generally fulfilled in practical applications. Moreover, for CBO methods there is recent work where such assumption about the initial datum is reduced to the absolute minimum [18, 19].

Remark 5 The choice of the parameter α_0 necessary in Theorem 2 may be affected by the dimensionality d of the optimization problem at hand. By establishing a quantitative nonasymptotic Laplace principle, this dependence is made explicit in the works [18, Proposition 18] and [19, Proposition 1], where the authors show that α_0 may be required to grow linearly in d, see [18, Remark 21].

Proof of Theorem 2 Let us define the time horizon

$$T := \inf \left\{ t \geq 0 : \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_t))] < \frac{1}{2}\mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_0))] \right\} \quad \text{with } \inf \emptyset = \infty.$$

Obviously, by continuity, $T > 0$. We claim that $T = \infty$, which we prove by contradiction in the following. Therefore, assume $T < \infty$. Then, for $t \in [0, T]$, we have

$$\frac{\lambda \gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2}\right) \frac{2e^{-\alpha \underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_t))]}$$

$$\geq \frac{\lambda \gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2}\right) \frac{4e^{-\alpha \underline{\mathcal{E}}}}{\mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_0))]} = \mu > 0,$$

where the positivity of μ is due to the well-preparation condition P1 of the initialization. Lemma 2 then provides an upper bound for the time derivative of the functional $\mathbb{E}[\mathcal{H}(t)]$,

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] &\leq -\frac{\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] - \mu\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] \\
&\leq -\min\left\{\frac{\gamma}{m}, \mu\right\}\left(\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] + \mathbb{E}[|\overline{V}_t|^2]\right) \qquad (2.22) \\
&\leq -\frac{2}{3}\frac{\min\{\gamma/m, \mu\}}{\left((\gamma/(2m))^2 + 1\right)}\mathbb{E}[\mathcal{H}(t)] =: -\chi\mathbb{E}[\mathcal{H}(t)],
\end{aligned}$$

where we made use of the upper bound of (2.7) as in Lemma 1 in the last inequality. The rate χ is defined implicitly and it is straightforward to check that $\chi < \gamma/m$. Grönwall's inequality implies

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)]\exp(-\chi t). \qquad (2.23)$$

Let us now investigate the evolution of the functional $\mathcal{X}(t) := \left(\mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_t))]\right)^2$. First note that

$$\dot{\mathcal{X}}(0) := \frac{d}{dt}\mathcal{X}(t)\Big|_{t=0} = -2\alpha\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))]\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))\langle\nabla\mathcal{E}(\overline{X}_0), \overline{V}_0\rangle].$$

Then, an application of Grönwall's inequality to Equation (2.15) from Lemma 3 and using the explicit bound of $\mathbb{E}[\mathcal{H}(t)]$ from (2.23) yields

$$
\begin{aligned}
\frac{d}{dt}\mathcal{X}(t) &\geq \dot{\mathcal{X}}(0)\exp\left(-\frac{\gamma}{m}t\right) \\
&\quad - 4\alpha e^{-2\alpha\underline{\mathcal{E}}}C_{\mathcal{E}}\left(1 + 2\frac{\lambda}{m}\left(\frac{2m}{\gamma}\right)^2\right)\int_0^t \mathbb{E}[\mathcal{H}(s)]\exp\left(-\frac{\gamma}{m}(t-s)\right)ds \\
&\geq \dot{\mathcal{X}}(0)\exp\left(-\frac{\gamma}{m}t\right) \\
&\quad - 4\alpha e^{-2\alpha\underline{\mathcal{E}}}C_{\mathcal{E}}\left(1 + 2\frac{\lambda}{m}\left(\frac{2m}{\gamma}\right)^2\right)\mathbb{E}[\mathcal{H}(0)]\frac{1}{\gamma/m - \chi}\left(\exp\left(-\chi t\right) - \exp\left(-\frac{\gamma}{m}t\right)\right) \\
&\geq \dot{\mathcal{X}}(0)\exp\left(-\frac{\gamma}{m}t\right) \\
&\quad - 4\alpha e^{-2\alpha\underline{\mathcal{E}}}C_{\mathcal{E}}\left(1 + 2\frac{\lambda}{m}\left(\frac{2m}{\gamma}\right)^2\right)\mathbb{E}[\mathcal{H}(0)]\frac{1}{\gamma/m - \chi}\exp\left(-\chi t\right),
\end{aligned}
$$

which, in turn, implies

$$
\mathcal{X}(t) \geq \mathcal{X}(0) - \frac{m}{\gamma}\left(-\dot{\mathcal{X}}(0)\right)_+ - \frac{4\alpha e^{-2\alpha\underline{\mathcal{E}}}C_{\mathcal{E}}}{\chi(\gamma/m - \chi)}\left(1 + 2\frac{\lambda}{m}\left(\frac{2m}{\gamma}\right)^2\right)\mathbb{E}[\mathcal{H}(0)]
$$

after discarding the positive parts. Recalling the definition of $\mathcal{X}$ and employing the second well-preparation condition P2, we can deduce that for all $t \in [0, T]$ it holds

$$
\begin{aligned}
\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))]\right)^2 &\geq \left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))]\right)^2 \\
&\quad - \frac{2m\alpha}{\gamma}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))]\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))\langle\nabla\mathcal{E}(\overline{X}_0), \overline{V}_0\rangle]\right)_+ \\
&\quad - \frac{4\alpha e^{-2\alpha\underline{\mathcal{E}}}C_{\mathcal{E}}}{\chi(\gamma/m - \chi)}\left(1 + 2\frac{\lambda}{m}\left(\frac{2m}{\gamma}\right)^2\right)\mathbb{E}[\mathcal{H}(0)] \\
&> \frac{1}{4}\left(\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))]\right)^2,
\end{aligned}
$$

which entails that there exists $\delta > 0$ such that $\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))] \geq \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))]$ $/2$ in $[T, T + \delta]$ as well, contradicting the definition of T and therefore showing the claim $T = \infty$.

As a consequence of (2.23) we have

$$
\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)]\exp(-\chi t) \quad \text{and} \quad \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_t))] \geq \frac{1}{2}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{X}_0))]
\tag{2.24}
$$

for all $t \geq 0$. In particular, by means of Lemma 1, for a suitable generic constant $C > 0$, we infer

$$\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] \leq C \exp(-\chi t), \quad \mathbb{E}[|\overline{V}_t|^2] \leq C \exp(-\chi t),$$
$$\text{and} \quad \mathbb{E}[|\overline{X}_t - x_\alpha(\rho_{X,t})|^2] \leq C \exp(-\chi t), \tag{2.25}$$

where the last inequality uses the fact (2.13). Moreover, with Jensen's inequality,

$$\left| \frac{d}{dt} \mathbb{E}[\overline{X}_t] \right| \leq \mathbb{E}[|\overline{V}_t|] \leq C \exp\left(-\chi t/2\right) \to 0 \quad \text{as } t \to \infty,$$

showing that $\mathbb{E}[\overline{X}_t] \to \widetilde{x}$ for some $\widetilde{x} \in \mathbb{R}^d$, which may depend on α and f_0. According to (2.25), $\overline{X}_t \to \widetilde{x}$ in mean-square and $x_\alpha(\rho_{X,t}) \to \widetilde{x}$, since

$$|x_\alpha(\rho_{X,t}) - \widetilde{x}|^2 \leq 3\mathbb{E}[|x_\alpha(\rho_{X,t}) - \overline{X}_t|^2]$$
$$+ 3\mathbb{E}[|\overline{X}_t - \mathbb{E}\overline{X}_t|^2] + 3|\mathbb{E}\overline{X}_t - \widetilde{x}|^2 \to 0 \quad \text{as } t \to \infty.$$

Eventually, by continuity of the objective function $\mathcal{E}$ and by the dominated convergence theorem, we conclude that $\mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_t))] \to e^{-\alpha \mathcal{E}(\widetilde{x})}$ as $t \to \infty$. Using this when taking the limit $t \to \infty$ in the second bound of (2.24) after applying the logarithm and multiplying both sides with $-1/\alpha$, we obtain

$$\mathcal{E}(\widetilde{x}) = \lim_{t \to \infty} \left(-\frac{1}{\alpha} \log \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_t))] \right) \leq -\frac{1}{\alpha} \log \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_0))] + \frac{1}{\alpha} \log 2. \tag{2.26}$$

The Laplace principle (1.5) on the other hand allows to choose $\widetilde{\alpha} \gg 1$ large enough such that for given $\varepsilon > 0$ it holds $-\frac{1}{\alpha} \log \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{X}_0))] - \underline{\mathcal{E}} < \varepsilon/2$ for any $\alpha \geq \widetilde{\alpha}$. Together with (2.26), this establishes $0 \leq \mathcal{E}(\widetilde{x}) - \underline{\mathcal{E}} \leq \varepsilon/2 + (\log 2)/\alpha \leq \varepsilon$ for $\alpha \geq \max\{\widetilde{\alpha}, (2 \log 2)/\varepsilon\}$. Finally, under the inverse continuity property A5 we additionally have $|\widetilde{x} - x^*| \leq (\mathcal{E}(\widetilde{x}) - \underline{\mathcal{E}})^\nu/\eta \leq \varepsilon^\nu/\eta$, concluding the proof. $\qquad \square$

3 Mean-Field Analysis of PSO with Memory Effects

Let us now turn back to the PSO dynamics (1.2) described in the introduction. The fundamental difference to what was analyzed in the preceding section is the presence of a personal memory of each particle, encoded through the additional state variable Y_t^i. It can be thought of as an approximation to the in-time best position $\arg \min_{\tau \leq t} \mathcal{E}(X_\tau^i)$ seen by the respective particle. Its dynamics is encoded in Equation (1.2b).

In this section we analyze (1.2) in the large particle limit, i.e., through its mean-field limit (1.8).

3.1 Well-Posedness of PSO with Memory Effects

Ensured by a sufficiently regularized implementation of the local best position $\overline{Y}$, we can show the well-posedness of the mean-field PSO dynamics (1.8), respectively, the associated Vlasov-Fokker-Planck equation (1.7). As regards uniqueness, it does not seem straightforward to extend the standard proof technique to the present setting due to the way the memory effects are implemented in (1.2b) and (1.8b). Therefore, in what follows, we merely prove existence of solutions and leave the development of a suitably modified proof technique for future research, see also Remark 8.

Theorem 3 *Let $\mathcal{E}$ satisfy Assumptions A1–A3. Moreover, let m, γ, λ_1, λ_2, σ_1, σ_2, α, β, θ, κ, $T > 0$. If $(\overline{X}_0, \overline{Y}_0, \overline{V}_0)$ is distributed according to $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$, then the nonlinear SDE (1.8) admits a strong solution up to time T with $\mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathbb{R}^d) \times \mathcal{C}([0,T], \mathbb{R}^d)$-valued paths. The associated law f has regularity $\mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^{3d}))$ and is a weak solution to the Vlasov-Fokker-Planck equation (1.7). In particular,*

$$\sup_{t\in[0,T]} \mathbb{E}[|\overline{X}_t|^4 + |\overline{Y}_t|^4 + |\overline{V}_t|^4] \leq \left(1 + 3\mathbb{E}[|\overline{X}_0|^4 + |\overline{Y}_0|^4 + |\overline{V}_0|^4]\right) e^{CT} \quad (3.1)$$

for some constant $C > 0$ depending only on m, γ, λ_1, λ_2, σ_1, σ_2, α, β, θ, κ, $c_{\mathcal{E}}$, R and $L_{\mathcal{E}}$.

Proof sketch The proof follows the steps taken in [8, Theorems 3.1, 3.2].
Step 1: For a given function $u \in \mathcal{C}([0,T], \mathbb{R}^d)$ and an initial measure $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$, according to standard SDE theory [2, Chapter 6], we can uniquely solve the auxiliary SDE

$$
\begin{aligned}
d\widetilde{X}_t &= \widetilde{V}_t \, dt, \\
d\widetilde{Y}_t &= \kappa\left(\widetilde{X}_t - \widetilde{Y}_t\right) S^{\beta,\theta}\left(\widetilde{X}_t, \widetilde{Y}_t\right) dt, \\
m \, d\widetilde{V}_t &= -\gamma \widetilde{V}_t \, dt + \lambda_1\left(\widetilde{Y}_t - \widetilde{X}_t\right) dt + \lambda_2\left(u_t - \widetilde{X}_t\right) dt + \sigma_1 D\left(\widetilde{Y}_t - \widetilde{X}_t\right) dB_t^1 \\
&\quad + \sigma_2 D\left(u_t - \widetilde{X}_t\right) dB_t^2,
\end{aligned}
$$

with initial condition $\left(\widetilde{X}_0, \widetilde{Y}_0, \widetilde{V}_0\right) \sim f_0$ as, due to the smoothness of $S^{\beta,\theta}$ and Assumptions A2 and A3, the coefficients are locally Lipschitz and have at most linear growth. This induces $\widetilde{f}_t = \mathrm{Law}\left(\widetilde{X}_t, \widetilde{Y}_t, \widetilde{V}_t\right)$. Moreover, the regularity of $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$ allows for a moment estimate of the form (3.1) and thus $\widetilde{f} \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^{3d}))$, see, e.g. [2, Chapter 7]. In what follows, $\widetilde{\rho}_Y$ denotes the spatial local best marginal of $\widetilde{f}$, i.e., $\widetilde{\rho}_Y(t, \cdot) = \iint_{\mathbb{R}^{2d}} d\widetilde{f}(t, x, \cdot, v)$.
Step 2: Let us now define, for some constant $C > 0$, the test function space

$$
\begin{aligned}
\mathcal{C}_*^2(\mathbb{R}^{3d}) := \Big\{ \phi \in \mathcal{C}^2(\mathbb{R}^{3d}) : |\nabla_v \phi| &\leq C \left(1 + |x| + |y| + |v|\right) \\
\text{and} \quad \sup_{k=1,\dots,d} \left\| \partial_{v_k v_k}^2 \phi \right\|_\infty &< \infty \Big\}.
\end{aligned}
\quad (3.2)
$$

For some $\phi \in \mathcal{C}_*^2(\mathbb{R}^{3d})$, by the Itô-Doeblin formula, we derive

$$
\begin{aligned}
d\phi = {} & \nabla_x \phi \cdot \widetilde{V}_t \, dt + \kappa \nabla_y \phi \cdot \left(\widetilde{X}_t - \widetilde{Y}_t\right) S^{\beta,\theta}\left(\widetilde{X}_t, \widetilde{Y}_t\right) dt \\
& + \nabla_v \phi \cdot \left(-\frac{\gamma}{m}\widetilde{V}_t + \frac{\lambda_1}{m}\left(\widetilde{Y}_t - \widetilde{X}_t\right) + \frac{\lambda_2}{m}\left(u_t - \widetilde{X}_t\right)\right) dt \\
& + \frac{1}{2}\sum_{k=1}^d \partial^2_{v_k v_k}\phi \left(\frac{\sigma_1^2}{m^2}\left(\widetilde{Y}_t - \widetilde{X}_t\right)^2_k + \frac{\sigma_2^2}{m^2}\left(u_t - \widetilde{X}_t\right)^2_k\right) dt \\
& + \nabla_v \phi \cdot \left(\frac{\sigma_1}{m}D\left(\widetilde{Y}_t - \widetilde{X}_t\right) dB_t^1 + \frac{\sigma_2}{m}D\left(u_t - \widetilde{X}_t\right) dB_t^2\right),
\end{aligned}
$$

where we mean $\phi\left(\widetilde{X}_t, \widetilde{Y}_t, \widetilde{V}_t\right)$ whenever we write ϕ. After taking the expectation, applying Fubini's theorem and observing that the stochastic integrals vanish due to the definition of the test function space $\mathcal{C}_*^2(\mathbb{R}^{3d})$ and the regularity (3.1), we observe that $\widetilde{f} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{3d}))$ satisfies the Vlasov-Fokker-Planck equation

$$
\begin{aligned}
\frac{d}{dt}\iiint_{\mathbb{R}^{3d}} \phi \, d\widetilde{f}_t = {} & \iiint_{\mathbb{R}^{3d}} v \cdot \nabla_x \phi \, d\widetilde{f}_t + \iiint_{\mathbb{R}^{3d}} \kappa(x-y)S^{\beta,\theta}(x,y) \cdot \nabla_y \phi \, d\widetilde{f}_t \\
& - \iiint_{\mathbb{R}^{3d}} \left(\frac{\gamma}{m}v + \frac{\lambda_1}{m}(x-y) + \frac{\lambda_2}{m}(x-u_t)\right) \cdot \nabla_v \phi \, d\widetilde{f}_t \\
& + \iiint_{\mathbb{R}^{3d}} \sum_{k=1}^d \left(\frac{\sigma_1^2}{2m^2}(x-y)^2_k + \frac{\sigma_2^2}{2m^2}(x-u_t)^2_k\right) \cdot \partial^2_{v_k v_k}\phi \, d\widetilde{f}_t.
\end{aligned}
\tag{3.3}
$$

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 the stability estimate $|y_\alpha(\widetilde{\rho}_{Y,t}) - y_\alpha(\widetilde{\rho}_{Y,s})|_2 \lesssim W_2(\widetilde{\rho}_{Y,t}, \widetilde{\rho}_{Y,s})$ for $\widetilde{\rho}_{Y,t}, \widetilde{\rho}_{Y,s} \in \mathcal{P}_4(\mathbb{R}^d)$, see, e.g., [8, Lemma 3.2], and the Hölder-1/2 continuity of the Wasserstein-2 distance $W_2(\widetilde{\rho}_{Y,t}, \widetilde{\rho}_{Y,s})$.

Step 4: Then, for $u = \vartheta \mathcal{T}u$ with $\vartheta \in [0, 1]$, there exists $\widetilde{f} \in \mathcal{C}([0, T], \mathcal{P}_4(\mathbb{R}^{3d}))$ satisfying (3.3) with marginal $\widetilde{\rho}_Y$ such that $u_t = \vartheta y_\alpha(\widetilde{\rho}_{Y,t})$. For such u, a uniform bound can be obtained as of Assumption A3. An application of the Leray-Schauder fixed point theorem provides a solution to (1.8). $\qquad\square$

3.2 Convergence of PSO with Memory Effects to a Global Minimizer

Analogously to Sect. 2.2 we define a functional $\mathcal{H}(t)$, which is analyzed in this section to eventually prove its exponential decay and thereby consensus formation at some $\widetilde{x}$ close to the global minimizer x^*. In addition to the requirements that the variance $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]$ in the position and the second-order moment of the

velocity $\mathbb{E}[|\overline{V}_t|^2]$ of the averaged particle vanish, we also expect that the particle's position $\overline{X}_t$ aligns with its personal best position $\overline{Y}_t$ over time, meaning that $\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$ decays to zero. This motivates the definition

$$
\begin{aligned}
\mathcal{H}(t) := {}& \left(\frac{\gamma}{2m}\right)^2 |\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + \frac{3}{2}|\overline{V}_t|^2 + \frac{1}{2}\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right)|\overline{X}_t - \overline{Y}_t|^2 \\
& + \frac{\gamma}{2m}\langle \overline{X}_t - \mathbb{E}[\overline{X}_t], \overline{V}_t\rangle + \frac{\gamma}{m}\langle \overline{X}_t - \overline{Y}_t, \overline{V}_t\rangle,
\end{aligned}
\tag{3.4}
$$

whose last two terms are required for technical reasons. Again, by the equivalence established in the following Lemma 4, proving the decay of the Lyapunov function $\mathbb{E}[\mathcal{H}(t)]$ directly entails the decay of $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2 + |\overline{X}_t - \overline{Y}_t|^2]$ with the same rate.

Lemma 4 *The functional $\mathcal{H}(t)$ is equivalent to $|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2 + |\overline{X}_t - \overline{Y}_t|^2$ in the sense that*

$$
\begin{aligned}
& \frac{1}{2}\left(\frac{\gamma}{2m}\right)^2 |\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + \frac{1}{2}|\overline{V}_t|^2 + \frac{3\lambda_1}{2m}|\overline{X}_t - \overline{Y}_t|^2 \le \mathcal{H}(t) \\
& \le \frac{5}{2}\left(\left(\frac{\gamma}{2m}\right)^2 + 1 + \frac{3\lambda_1}{m} + \frac{2\gamma^2}{m^2}\right)\left(|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2 + |\overline{X}_t - \overline{Y}_t|^2\right).
\end{aligned}
\tag{3.5}
$$

We now derive an evolution inequality of the Lyapunov function $\mathbb{E}[\mathcal{H}(t)]$.

Lemma 5 *Let $\mathcal{E}$ satisfy Assumptions A1–A3 and let $(\overline{X}_t, \overline{Y}_t, \overline{V}_t)_{t\ge 0}$ be a solution to the nonlinear SDE (1.8). Then $\mathbb{E}[\mathcal{H}(t)]$ with $\mathcal{H}$ as defined in (3.4) satisfies*

$$
\begin{aligned}
& \frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \\
& \le -\frac{\gamma}{2m}\mathbb{E}[|\overline{V}_t|^2] \\
& \quad - \left(\frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2} - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2}\right)\frac{6e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))]}\right)\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] \\
& \quad - \left(\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} + \kappa\theta\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} - \frac{3\sigma_1^2}{2m^2}\right. \\
& \quad \left. - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2}\right) - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2}\right)\frac{12e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))]}\right)\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2].
\end{aligned}
\tag{3.6}
$$

Proof Let us write $\delta\overline{X}_t := \overline{X}_t - \mathbb{E}[\overline{X}_t]$ for short and note that the integration by parts formula gives

$$
\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] = 2\mathbb{E}[\langle \delta\overline{X}_t, \overline{V}_t\rangle].
\tag{3.7}
$$

Observe that the stochastic integrals have vanishing expectations as a consequence of [39, Theorem 3.2.1(iii), Definition 3.1.4(iii)] combined with the regularity $f \in \mathcal{C}([0,T], \mathcal{P}_4(\mathbb{R}^{2d}))$ obtained in Theorem 3. An application of the Itô-Doeblin formula and Young's inequality yields

$$
\begin{aligned}
\frac{d}{dt}\mathbb{E}[|\overline{V}_t|^2] &= -\frac{2\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] + \frac{2\lambda_1}{m}\mathbb{E}[\langle \overline{V}_t, \overline{Y}_t - \overline{X}_t\rangle] \\
&\quad + \frac{2\lambda_2}{m}\mathbb{E}[\langle \overline{V}_t, y_\alpha(\rho_{Y,t}) - \overline{X}_t\rangle] + \frac{\sigma_1^2}{m^2}\mathbb{E}[|\overline{Y}_t - \overline{X}_t|^2] \\
&\quad + \frac{\sigma_2^2}{m^2}\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{X}_t|^2] \\
&\leq -\left(\frac{2\gamma}{m} - \frac{\lambda_2}{\varepsilon m}\right)\mathbb{E}[|\overline{V}_t|^2] + \frac{\sigma_1^2}{m^2}\mathbb{E}[|\overline{Y}_t - \overline{X}_t|^2] \\
&\quad + \left(\frac{\varepsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right)\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{X}_t|^2] \\
&\quad - \frac{2\lambda_1}{m}\mathbb{E}[\langle \overline{V}_t, \overline{X}_t - \overline{Y}_t\rangle], \quad \forall \varepsilon > 0.
\end{aligned}
\tag{3.8}
$$

Again by employing the Itô-Doeblin formula we obtain

$$
\begin{aligned}
\frac{d}{dt}\mathbb{E}[\langle \delta\overline{X}_t, \overline{V}_t\rangle] &= \mathbb{E}[|\overline{V}_t|^2] - \left(\mathbb{E}[\overline{V}_t]\right)^2 - \frac{\gamma}{m}\mathbb{E}[\langle \delta\overline{X}_t, \overline{V}_t\rangle] + \frac{\lambda_1}{m}\mathbb{E}[\langle \delta\overline{X}_t, \overline{Y}_t - \overline{X}_t\rangle] \\
&\quad + \frac{\lambda_2}{m}\mathbb{E}[\langle \delta\overline{X}_t, y_\alpha(\rho_{Y,t}) - \overline{X}_t\rangle] \\
&\leq \mathbb{E}[|\overline{V}_t|^2] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] \\
&\quad + \frac{\lambda_1}{m}\mathbb{E}[\langle \delta\overline{X}_t, \left(\overline{Y}_t - y_\alpha(\rho_{Y,t})\right) - \left(\overline{X}_t - \mathbb{E}[\overline{X}_t]\right)\rangle] \\
&\quad + \frac{\lambda_2}{m}\mathbb{E}[\langle \delta\overline{X}_t, \mathbb{E}[\overline{X}_t] - \overline{X}_t\rangle] \\
&= \mathbb{E}[|\overline{V}_t|^2] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] - \frac{\lambda_1 + \lambda_2}{m}\mathbb{E}[|\delta\overline{X}_t|^2] \\
&\quad + \frac{\lambda_1}{m}\mathbb{E}[\langle \delta\overline{X}_t, \overline{Y}_t - y_\alpha(\rho_{Y,t})\rangle] \\
&\leq \mathbb{E}[|\overline{V}_t|^2] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] - \frac{\lambda_1 + 2\lambda_2}{2m}\mathbb{E}[|\delta\overline{X}_t|^2] \\
&\quad + \frac{\lambda_1}{2m}\mathbb{E}[|\overline{Y}_t - y_\alpha(\rho_{Y,t})|^2],
\end{aligned}
$$

where, for the second line, we used the identity (3.7) and that $\mathbb{E}[\langle \delta\overline{X}_t, \mathbf{C}\rangle] = 0$, whenever $\mathbf{C} \in \mathbb{R}^d$ is constant, allowing to expand the expression in the way done. We

now rearrange the previous inequality to get

$$\frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_t|^2] + \frac{d}{dt}\mathbb{E}[\langle\delta\overline{X}_t, \overline{V}_t\rangle] \leq \mathbb{E}[|\overline{V}_t|^2] - \frac{\lambda_1 + 2\lambda_2}{2m}\mathbb{E}[|\delta\overline{X}_t|^2]$$
$$+\frac{\lambda_1}{2m}\mathbb{E}[|\overline{Y}_t - y_\alpha(\rho_{Y,t})|^2]. \tag{3.9}$$

Next, using the Itô-Doeblin formula, we compute

$$\frac{d}{dt}\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] = 2\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, \overline{V}_t - \kappa\left(\overline{X}_t - \overline{Y}_t\right)S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t)\rangle]$$
$$\leq 2\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, \overline{V}_t\rangle] - 2\kappa\theta\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2], \tag{3.10}$$

where the last step follows from the fact that $\theta < S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t) < 2 + \theta < 4$. And lastly, the Itô-Doeblin formula and Young's inequality allow to bound

$$\frac{d}{dt}\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, \overline{V}_t\rangle]$$
$$= -\frac{\gamma}{m}\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, \overline{V}_t\rangle] - \frac{\lambda_1 + \lambda_2}{m}\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] + \frac{\lambda_2}{m}\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, y_\alpha(\rho_{Y,t}) - \overline{Y}_t\rangle]$$
$$+ \mathbb{E}[\langle\overline{V}_t - \kappa\left(\overline{X}_t - \overline{Y}_t\right)S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t), \overline{V}_t\rangle]$$
$$\leq -\frac{\gamma}{m}\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, \overline{V}_t\rangle] - \frac{\lambda_1 + \lambda_2}{m}\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] + \frac{\lambda_2^2}{2m\lambda_1}\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$$
$$+ \frac{\lambda_1}{2m}\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{Y}_t|^2]$$
$$+ \mathbb{E}[|\overline{V}_t|^2] + \frac{1}{2}\mathbb{E}[|\overline{V}_t|^2] + 8\kappa^2\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$$
$$= -\left(\frac{\lambda_1 + \lambda_2}{m} - 8\kappa^2 - \frac{\lambda_2^2}{2m\lambda_1}\right)\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] + \frac{3}{2}\mathbb{E}[|\overline{V}_t|^2] + \frac{\lambda_1}{2m}\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{Y}_t|^2]$$
$$- \frac{\gamma}{m}\mathbb{E}[\langle\overline{X}_t - \overline{Y}_t, \overline{V}_t\rangle]. \tag{3.11}$$

We now collect the bounds (3.8), (3.9), (3.10), and (3.11) to show

$$\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)]$$
$$\leq -\left(\frac{3\gamma}{m} - \frac{3\lambda_2}{2\varepsilon m} - \frac{\gamma}{2m} - \frac{3\gamma}{2m}\right)\mathbb{E}[|\overline{V}_t|^2] - \frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2}\mathbb{E}[|\delta\overline{X}_t|^2]$$
$$-\left(\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} + \kappa\theta\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) - \frac{3\sigma_1^2}{2m^2}\right)\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$$
$$+\frac{3}{2}\left(\frac{\varepsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right)\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{X}_t|^2] + \frac{3\lambda_1\gamma}{(2m)^2}\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{Y}_t|^2]$$
$$\leq -\left(\frac{\gamma}{m} - \frac{3\lambda_2}{2\varepsilon m}\right)\mathbb{E}[|\overline{V}_t|^2] - \frac{(\lambda_1 + 2\lambda_2)\gamma}{(2m)^2}\mathbb{E}[|\delta\overline{X}_t|^2]$$
$$-\left(\frac{(\lambda_1 + \lambda_2)\gamma}{m^2} - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} + \kappa\theta\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) - \frac{3\sigma_1^2}{2m^2} - 3\left(\frac{\varepsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right)\right)\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$$

$$+ \left(3\left(\frac{\varepsilon\lambda_2}{m} + \frac{\sigma_2^2}{m^2}\right) + \frac{3\lambda_1\gamma}{(2m)^2}\right) \mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{Y}_t|^2].$$

Recalling the computation (2.13) yields the bound

$$\mathbb{E}[|\overline{Y}_t - y_\alpha(\rho_{Y,t})|^2] \leq 2e^{-\alpha\mathcal{E}} \frac{\mathbb{E}[|\delta\overline{Y}_t|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))]}$$

$$\leq 2e^{-\alpha\mathcal{E}} \frac{6\mathbb{E}[|\overline{Y}_t - \overline{X}_t|^2] + 3\mathbb{E}[|\delta\overline{X}_t|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))]}, \qquad (3.12)$$

where we inserted $\pm\overline{X}_t$ and $\pm\mathbb{E}[\overline{X}_t]$ in the second step and used that $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ as well as Jensen's inequality. Combining the last two bounds and choosing $\varepsilon = (3\lambda_2)/\gamma$ we obtain (3.6) as desired. $\qquad\qquad\square$

Remark 6 The exponential decay of $\mathbb{E}[\mathcal{H}(t)]$ it obtained by choosing the parameters of PSO in a manner which ensures the negativity of the prefactors of $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]$ and $\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$ in Inequality (3.6). This may be achieved by choosing for any fixed time t, given α and arbitrary $\theta, \sigma_1, \sigma_2, \gamma > 0$,

$$\lambda_1 > \frac{3\sigma_1^2}{2\gamma}, \ \lambda_2 > 6\max\left\{\frac{D_t^Y\lambda_1}{4}, \frac{(1 + D_t^Y)\sigma_2^2}{\gamma}\right\}, \ \kappa > \frac{3\lambda_2^2(1 + D_t^Y)}{\gamma\theta\lambda_1},$$

$$\text{and } m < \min\left\{\frac{\gamma\theta}{16\kappa}, \frac{\lambda_1\gamma^2}{18D_t^Y\lambda_2^2}\right\},$$

where we abbreviate $D_t^Y = 12e^{-\alpha\mathcal{E}}/\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))]$.

In our main theorem on convergence of the PSO dynamics with memory mechanisms to the global minimizer x^* we again ensure that the parameter can be chosen once at initialization time.

Theorem 4 *Let $\mathcal{E}$ satisfy Assumptions A1–A4 and let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be a solution to the nonlinear SDE (1.8). Moreover, let us assume the well-preparation of the initial datum $\overline{X}_0$ and $\overline{V}_0$ in the sense that*

P1 $\mu_1 > 0$ with

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

P2 $\mu_2 > 0$ with

$$\mu_2 := \frac{(\lambda_1 + \lambda_2)\gamma}{m^2} + \kappa\theta\left(\frac{3\lambda_1}{m} + \frac{\gamma^2}{m^2}\right) - \frac{8\kappa^2\gamma}{m} - \frac{\lambda_2^2\gamma}{2m^2\lambda_1} - \frac{3\sigma_1^2}{2m^2}$$
$$- \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2}\right) - \left(\frac{9\lambda_2^2}{\gamma m} + \frac{3\sigma_2^2}{m^2} + \frac{3\lambda_1\gamma}{(2m)^2}\right) \frac{24e^{-\alpha\mathcal{E}}}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_0))]},$$

P3 it holds

$$\left(\frac{\alpha \kappa m}{\lambda_1 \chi} \left(C_{\mathcal{E}} + 2\alpha^2 \right) + \frac{24 C_{\mathcal{E}}^2 \kappa}{\alpha \chi^3} \right) \frac{\mathbb{E}[\mathcal{H}(0)]}{\mathbb{E}[\exp(-\alpha(\mathcal{E}(\overline{Y}_0) - \underline{\mathcal{E}}))]}$$
$$+ \frac{6\kappa}{\alpha \chi} \frac{\mathbb{E}[|\nabla \mathcal{E}(\overline{X}_0)|^2]}{\mathbb{E}[\exp(-\alpha(\mathcal{E}(\overline{Y}_0) - \underline{\mathcal{E}}))]} < \frac{3}{32}$$

where

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

Then $\mathbb{E}[\mathcal{H}(t)]$ with $\mathcal{H}$ as defined in Equation (3.4) converges exponentially fast with rate χ to 0 as $t \to \infty$. Moreover, there exists some $\widetilde{x}$, which may depend on α and f_0, such that $\mathbb{E}[\overline{X}_t] \to \widetilde{x}$ and $y_\alpha(\rho_{Y,t}) \to \widetilde{x}$ exponentially fast with rate $\chi/2$ as $t \to \infty$. Eventually, for any given accuracy $\varepsilon > 0$, there exists $\alpha_0 > 0$, which may depend on the dimension d, such that for all $\alpha > \alpha_0$, $\widetilde{x}$ satisfies

$$\mathcal{E}(\widetilde{x}) - \underline{\mathcal{E}} \leq \varepsilon.$$

If $\mathcal{E}$ additionally satisfies Assumption A5, we additionally have $|\widetilde{x} - x^*| \leq \varepsilon^\nu/\eta$.

Remark 7 By replacing D_t^Y with $2D_0^Y$ in the parameter choices of Remark 6, the well-preparation of the parameters as in Conditions P1 and P2 can be ensured.

In analogy to Remark 4, Condition P3 guarantees the well-preparation of the initialization.

Proof of Theorem 4 Let us define the time horizon

$$T := \inf \left\{ t \geq 0 : \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{Y}_t))] < \frac{1}{2} \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{Y}_0))] \right\} \quad \text{with } \inf \emptyset = \infty.$$

Obviously, by continuity, $T > 0$. We claim that $T = \infty$, which we prove by contradiction in the following. Therefore, assume $T < \infty$. Then, for $t \in [0, T]$, noting that $\mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{Y}_t))] \geq \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{Y}_0))]/2$, we observe that the prefactors of $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2]$ and $\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$ in Lemma 5 are upper bounded by $-\mu_1$ and $-\mu_2$, respectively. Lemma 5 then provides an upper bound for the time derivative of the functional $\mathbb{E}[\mathcal{H}(t)]$,

$$\frac{d}{dt} \mathbb{E}[\mathcal{H}(t)] \leq -\frac{\gamma}{2m} \mathbb{E}[|\overline{V}_t|^2] - \mu_1 \mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] - \mu_2 \mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2]$$
$$\leq -\min \left\{ \frac{\gamma}{2m}, \mu_1, \mu_2 \right\} \left(\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] + \mathbb{E}[|\overline{V}_t|^2] + \mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] \right) \quad (3.13)$$
$$\leq -\frac{2}{5} \frac{\min\{\gamma/(2m), \mu_1, \mu_2\}}{\left((\gamma/(2m))^2 + 1 + 3\lambda_1/m + 2\gamma^2/m^2 \right)} \mathbb{E}[\mathcal{H}(t)] =: -\chi \mathbb{E}[\mathcal{H}(t)],$$

where we made use of the upper bound of (3.5) as in Lemma 4 in the last inequality. The rate χ is defined implicitly and it is straightforward to check that $0 < \chi < \gamma/m$,

where the positivity of χ follows from the well-preparation conditions P1 and P2 of the initialization. Grönwall's inequality implies

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t). \qquad (3.14)$$

We now investigate the evolution of the functional $\mathcal{Y}(t) := \mathbb{E}[\exp(-\alpha \mathcal{E}(\overline{Y}_t))]$. The Itô-Doeblin formula yields

$$
\begin{aligned}
\frac{d}{dt}\mathcal{Y}(t) &= -\alpha\kappa\,\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))\,\langle \nabla\mathcal{E}(\overline{Y}_t), (\overline{X}_t - \overline{Y}_t)S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t)\rangle] \\
&= -\alpha\kappa\,\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))\,\langle \nabla\mathcal{E}(\overline{Y}_t) - \nabla\mathcal{E}(\overline{X}_t), (\overline{X}_t - \overline{Y}_t)S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t)\rangle] \\
&\quad - \alpha\kappa\,\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))\,\langle \nabla\mathcal{E}(\overline{X}_t), (\overline{X}_t - \overline{Y}_t)S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t)\rangle] \\
&\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}C_\mathcal{E}\,\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] - 4\alpha\kappa e^{-\alpha\mathcal{E}}\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_t)||\overline{X}_t - \overline{Y}_t|],
\end{aligned}
$$
$$(3.15)$$

where the last step follows from Cauchy-Schwarz inequality and uses Assumption A4 and $S^{\beta,\theta}(\overline{X}_t, \overline{Y}_t) < 4$. Now firstly notice that $\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_t)||\overline{X}_t - \overline{Y}_t|] \leq e^{(\chi/2)t}\alpha^2\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] + e^{-(\chi/2)t}/\alpha^2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_t)|^2]$ by Young's inequality. Secondly, using again Assumption A4 in the first inequality, we have

$$
\begin{aligned}
\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_t)|^2] &= \mathbb{E}\left[\left|\nabla\mathcal{E}(\overline{X}_0) + \int_0^t \nabla^2\mathcal{E}(\overline{X}_s)\overline{V}_s\,ds\right|^2\right] \\
&\leq 2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2] + 2C_\mathcal{E}^2 t\int_0^t \mathbb{E}[|\overline{V}_s|^2]\,ds \\
&\leq 2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2] + 4C_\mathcal{E}^2 t\int_0^t \mathbb{E}[\mathcal{H}(s)]\,ds \\
&\leq 2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2] + 4C_\mathcal{E}^2 t\mathbb{E}[\mathcal{H}(0)]\int_0^t \exp(-\chi s)\,ds \\
&= 2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2] + 4C_\mathcal{E}^2 t\mathbb{E}[\mathcal{H}(0)]\frac{1}{\chi}\left(1 - \exp(-\chi t)\right),
\end{aligned}
$$

where the next-to-last step uses the explicit bound in (3.14). Using the two latter observations together with the fact that $\mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] \leq 2m/(3\lambda_1)\mathbb{E}[\mathcal{H}(t)]$ we can continue (3.15) as follows

$$
\begin{aligned}
\frac{d}{dt}\mathcal{Y}(t) &\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}\left(C_\mathcal{E} + \exp\left(\frac{\chi}{2}t\right)\alpha^2\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(t)] \\
&\quad - \frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\exp\left(-\frac{\chi}{2}t\right)\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_t)|^2] \\
&\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}\left(C_\mathcal{E} + \exp\left(\frac{\chi}{2}t\right)\alpha^2\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(0)]\exp(-\chi t) \\
&\quad - \frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\exp\left(-\frac{\chi}{2}t\right)\left(2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2] + 4C_\mathcal{E}^2 t\mathbb{E}[\mathcal{H}(0)]\frac{1}{\chi}\left(1 - \exp(-\chi t)\right)\right)
\end{aligned}
$$

$$\geq -4\alpha\kappa e^{-\alpha\mathcal{E}}\left(C_{\mathcal{E}}\exp\left(-\chi t\right) + \exp\left(-\frac{\chi}{2}t\right)\alpha^2\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(0)]$$

$$-\frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\exp\left(-\frac{\chi}{2}t\right)\left(2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2] + \frac{4C_{\mathcal{E}}^2 t}{\chi}\mathbb{E}[\mathcal{H}(0)]\right). \tag{3.16}$$

By integrating (3.16) we obtain for all $t \in [0, T]$

$$\mathcal{Y}(t) \geq \mathcal{Y}(0) - 4\alpha\kappa e^{-\alpha\mathcal{E}}\left(\frac{C_{\mathcal{E}}}{\chi} + \frac{2\alpha^2}{\chi}\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(0)]$$

$$-\frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\left(2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2]\frac{2}{\chi} + \frac{16C_{\mathcal{E}}^2}{\chi^3}\mathbb{E}[\mathcal{H}(0)]\right).$$

Recalling the definition of $\mathcal{Y}$ and employing Condition P3, we can deduce that for all $t \in [0, T]$ it holds

$$\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))] \geq \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_0))] - 4\alpha\kappa e^{-\alpha\mathcal{E}}\left(\frac{C_{\mathcal{E}}}{\chi} + \frac{2\alpha^2}{\chi}\right)\frac{2m}{3\lambda_1}\mathbb{E}[\mathcal{H}(0)]$$

$$-\frac{4}{\alpha}\kappa e^{-\alpha\mathcal{E}}\left(2\mathbb{E}[|\nabla\mathcal{E}(\overline{X}_0)|^2]\frac{2}{\chi} + \frac{16C_{\mathcal{E}}^2}{\chi^3}\mathbb{E}[\mathcal{H}(0)]\right)$$

$$> \frac{3}{4}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_0))],$$

which entails that there exists $\delta > 0$ such that $\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))] \geq \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_0))]/2$ in $[T, T+\delta]$ as well, contradicting the definition of T and therefore showing the claim $T = \infty$.

As a consequence of (3.14) we have

$$\mathbb{E}[\mathcal{H}(t)] \leq \mathbb{E}[\mathcal{H}(0)]\exp(-\chi t) \quad \text{and} \quad \mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_t))] \geq \frac{1}{2}\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_0))] \tag{3.17}$$

for all $t \geq 0$. In particular, by means of Lemma 4, for a suitable generic constant $C > 0$, we infer

$$\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] \leq C\exp(-\chi t), \quad \mathbb{E}[|\overline{V}_t|^2] \leq C\exp(-\chi t),$$
$$\text{and} \quad \mathbb{E}[|\overline{X}_t - \overline{Y}_t|^2] \leq C\exp(-\chi t). \tag{3.18}$$

Moreover, with Jensen's inequality,

$$\left|\frac{d}{dt}\mathbb{E}[\overline{X}_t]\right| \leq \mathbb{E}[|\overline{V}_t|] \leq C \exp\left(-\chi t/2\right) \to 0 \quad \text{as } t \to \infty,$$

showing that $\mathbb{E}[\overline{X}_t] \to \widetilde{x}$ for some $\widetilde{x} \in \mathbb{R}^d$, which may depend on α and f_0. According to (3.18), $\overline{X}_t \to \widetilde{x}$ as well as $\overline{Y}_t \to \widetilde{x}$ in mean-square. Moreover, by reusing the inequality (3.12) we get

$$\mathbb{E}[|\overline{Y}_t - y_\alpha(\rho_{Y,t})|^2] \leq 4e^{-\alpha\underline{\mathcal{E}}}\frac{6\mathbb{E}[|\overline{Y}_t - \overline{X}_t|^2] + 3\mathbb{E}[|\overline{X}_t - \mathbb{E}\overline{X}_t|^2]}{\mathbb{E}[\exp(-\alpha\mathcal{E}(\overline{Y}_0))]} \leq C\exp(-\chi t)$$

showing $y_\alpha(\rho_{Y,t}) \to \widetilde{x}$, since

$$|y_\alpha(\rho_{Y,t}) - \widetilde{x}|^2 \leq 4\mathbb{E}[|y_\alpha(\rho_{Y,t}) - \overline{Y}_t|^2] + 4\mathbb{E}[|\overline{Y}_t - \overline{X}_t|^2]$$
$$+ 4\mathbb{E}[|\overline{X}_t - \mathbb{E}\overline{X}_t|^2] + 4|\mathbb{E}\overline{X}_t - \widetilde{x}|^2 \to 0 \quad \text{as } t \to \infty.$$

The remainder of the proof follows the lines of the proof of Theorem 2, replacing merely $\overline{X}_t$ with $\overline{Y}_t$. □

4 A Holistic Convergence Statement of PSO Without Memory Effects

In Sects. 2 and 3 we analyzed the macroscopic behavior of PSO without and with memory effects in the mean-field regime. For this purpose we introduced the with (1.2) and (2.1) associated self-consistent mono-particle processes (1.8) and (2.3), for which we then established convergence guarantees under the in Theorems 2 and 4 specified assumptions. However, in order to be able to infer therefrom the optimization capabilities of the numerically implemented PSO method, a quantitative estimate on the approximation quality of the interacting particle system by the corresponding mean-field dynamics is necessary.

4.1 On the Mean-Field Approximation of PSO Without Memory Effects

The following theorem provides a probabilistic quantitative estimate on the mean-field approximation for PSO without memory effects. Notably, the result does not suffer from the curse of dimensionality.

Theorem 5 *Let $T > 0$, $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$ and let $N \in \mathbb{N}$ be fixed. Moreover, let $\mathcal{E}$ obey Assumptions A1–A4. We denote by $\left((X_t^i, V_t^i)_{t\geq 0}\right)_{i=1,\dots,N}$ the solution to system (2.1) and let $\left((\overline{X}_t^i, \overline{V}_t^i)_{t\geq 0}\right)_{i=1,\dots,N}$ be N independent copies of the solution to the mean-field*

dynamics (2.3). Then it holds

$$
\mathbb{P}(\Omega_M) = \mathbb{P}\left(\sup_{t\in[0,T]}\left[\frac{1}{N}\sum_{i=1}^{N}\max\left\{|X_t^i|^4 + |V_t^i|^4, |\overline{X}_t^i|^4 + |\overline{V}_t^i|^4\right\}\right] \le M\right)
$$
$$
\ge 1 - \frac{2K}{M}, \tag{4.1}
$$

where $K = K(\gamma/m, \lambda/m, \sigma/m, T, \mathcal{E})$ is a constant, which is in particular independent of N and d.

Furthermore, if the processes share the initial data as well as the Brownian motion paths $(B_t^i)_{t\ge 0}$ for all $i = 1, \ldots, N$, then we have a probabilistic mean-field approximation of the form

$$
\max_{i=1,\ldots,N}\sup_{t\in[0,T]}\mathbb{E}\left[|X_t^i - \overline{X}_t^i|^2 + |V_t^i - \overline{V}_t^i|^2 \,\Big|\, \Omega_M\right] \le C_{\text{MFA}}N^{-1} \tag{4.2}
$$

with a constant $C_{\text{MFA}} = C_{\text{MFA}}(\alpha, \gamma/m, \lambda/m, \sigma/m, T, \mathcal{E}, K, M)$, which is in particular independent of N and d.

Proof The proof is based on the arguments of [18, Section 3.3] about the mean-field approximation of CBO. First we compute a bound for $\mathbb{E}[\sup_{t\in[0,T]}\frac{1}{N}\sum_{i=1}^{N}\max\{|X_t^i|^4 + |V_t^i|^4, |\overline{X}_t^i|^4 + |\overline{V}_t^i|^4\}]$, which is then used to derive a mean-field approximation for PSO conditioned on the set Ω_M of uniformly bounded processes.

Step 1: Using standard inequalities and Jensen's inequality allows to derive the bound

$$
\mathbb{E}\left[\sup_{t\in[0,T]}|X_t^i|^4\right] \lesssim \mathbb{E}[|X_0^i|^4] + \mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t V_s^i\,ds\right|^4\right]
$$
$$
\le C\left(\mathbb{E}[|X_0^i|^4] + \mathbb{E}\left[\int_0^T |V_s^i|^4\,ds\right]\right) \tag{4.3}
$$

with $C = C(T)$. For the velocities V_t^i we first note that

$$
\mathbb{E}\left[\sup_{t\in[0,T]}|V_t^i|^4\right] \lesssim \mathbb{E}[|V_0^i|^4] + \left(\frac{\gamma}{m}\right)^4\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t V_s^i\,ds\right|^4\right]
$$
$$
+ \left(\frac{\lambda}{m}\right)^4\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t \left(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i\right)ds\right|^4\right] \tag{4.4}
$$
$$
+ \left(\frac{\sigma}{m}\right)^4\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t D\left(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i\right)dB_s^i\right|^4\right].
$$

While the two middle terms on the right-hand side of (4.4) can be controlled as before by applying Jensen's inequality, the last term is treated as follows. Since

$\int_0^t D\big(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i\big)dB_s^i$ is a martingale we can apply the Burkholder-Davis-Gundy inequality [47, Chapter IV, Theorem 4.1], which gives

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\int_0^t D\big(x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i\big)dB_s^i\right|^4\right] \lesssim \sup_{t\in[0,T]}\mathbb{E}\left[\left(\int_0^t \left|x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i\right|^2 ds\right)^2\right]$$
$$\leq C\mathbb{E}\left[\int_0^T \left|x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i\right|^4 ds\right],$$

$$(4.5)$$

where the latter step is again due to Jensen's inequality and with a constant $C = C(T)$. Utilizing these bounds allows to continue the inequality in (4.4) and to obtain

$$\mathbb{E}\left[\sup_{t\in[0,T]}|V_t^i|^4\right] \leq C\left(\mathbb{E}[|V_0^i|^4] + \mathbb{E}\left[\int_0^T |X_s^i|^4 + |x_\alpha(\widehat{\rho}_{X,s}^N)|^4 + |V_s^i|^4\, ds\right]\right) \quad (4.6)$$

with $C = C(\gamma/m, \lambda/m, \sigma/m, T)$. Since according to [8, Lemma 3.3] it holds

$$|x_\alpha(\widehat{\rho}_{X,s}^N)|^2 \leq \int |x|^2 \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\widehat{\rho}_{X,s}^N)}}\, d\widehat{\rho}_{X,s}^N(x) \leq b_1 + b_2 \int |x|^2\, d\widehat{\rho}_{X,s}^N(y)$$
$$= b_1 + b_2 \frac{1}{N}\sum_{i=1}^N |X_s^i|^2$$

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

$$b_1 = R^2 + b_2^2 \quad \text{and} \quad b_2 = \frac{2L_{\mathcal{E}}\max\{1, |x^*|^2\}}{c_{\mathcal{E}}}\left(1 + \frac{1}{\alpha c_{\mathcal{E}} R^2}\right)$$

in the case that $\mathcal{E}$ satisfies the coercivity assumption A3, we eventually obtain the upper bound

$$\mathbb{E}\left[\sup_{t\in[0,T]}|V_t^i|^4\right]$$
$$\leq C\left(1 + \mathbb{E}[|V_0^i|^4] + \mathbb{E}\left[\int_0^T |X_s^i|^4 + \frac{1}{N}\sum_{j=1}^N |X_s^j|^4 + |V_s^i|^4\, ds\right]\right)$$

$$(4.7)$$

with $C = C(\gamma/m, \lambda/m, \sigma/m, T, b_1, b_2)$. Adding up (4.3) and (4.7) yields

$$
\mathbb{E}\left[\sup_{t \in [0,T]} |X_t^i|^4 + |V_t^i|^4\right] \leq C \left(1 + \mathbb{E}[|X_0^i|^4 + |V_0^i|^4]\right.
$$
$$
\left. + \mathbb{E}\left[\int_0^T |X_s^i|^4 + \frac{1}{N}\sum_{j=1}^N |X_s^j|^2 + |V_s^i|^4 \, ds\right]\right),
\tag{4.8}
$$

which, averaged over i, allows to derive the bound

$$
\mathbb{E}\left[\sup_{t \in [0,T]} \frac{1}{N}\sum_{i=1}^N \left(|X_t^i|^4 + |V_t^i|^4\right)\right]
$$
$$
\leq C \left(1 + \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left(|X_0^i|^4 + |V_0^i|^4\right)\right] + \int_0^T \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left(|X_s^i|^4 + |V_s^i|^4\right)\right] ds\right).
\tag{4.9}
$$

An application of Grönwall's inequality ensures that $\mathbb{E}\sup_{t \in [0,T]}\left[\frac{1}{N}\sum_{i=1}^N\left(|X_t^i|^4 + |V_t^i|^4\right)\right]$ is bounded independently of N by some constant $K = K(\gamma/m, \lambda/m, \sigma/m, T, b_1, b_2)$. Note, that the constant K does in particular not depend on N or d. With identical arguments for the processes $(\overline{X}_t^i, \overline{V}_t^i)$ an analogous bound can be obtained for $\mathbb{E}\left[\sup_{t \in [0,T]} \frac{1}{N}\sum_{i=1}^N\left(|\overline{X}_t^i|^4 + |\overline{V}_t^i|^4\right)\right]$. The first claim of the statement now follows from Markov's inequality.

Step 2: We define the cutoff function

$$
I_M(t) = \begin{cases} 1, & \text{if } \frac{1}{N}\sum_{i=1}^N \max\left\{|X_s^i|^4 + |V_s^i|^4, |\overline{X}_s^i|^4 + |\overline{V}_s^i|^4\right\} \leq M \text{ for all } s \in [0,t], \\ 0, & \text{else,} \end{cases}
\tag{4.10}
$$

which is a random variable adapted to the natural filtration and satisfying $\mathbb{1}_{\Omega_M} \leq I_M(t)$ pointwise for all $t \in [0,T]$ as well as $I_M(t) = I_M(t)I_M(s)$ for all $s \in [0,t]$. Firstly, for the positions, by using standard inequalities and Jensen's inequality, we obtain the bound

$$
\mathbb{E}[|X_t^i - \overline{X}_t^i|^2 I_M(t)] \lesssim \mathbb{E}[|X_0^i - \overline{X}_0^i|^2] + \mathbb{E}\left[\left|\int_0^t (V_s^i - \overline{V}_s^i)I_M(s)\,ds\right|^2\right]
$$
$$
\leq C\left(\mathbb{E}[|X_0^i - \overline{X}_0^i|^2] + \int_0^t \mathbb{E}\left[|V_s^i - \overline{V}_s^i|^2 I_M(s)\right] ds\right)
\tag{4.11}
$$

with $C = C(T)$. Secondly, for the velocities we have

$$
\mathbb{E}[|V_t^i - \overline{V}_t^i|^2 I_M(t)]
$$

$$
\lesssim \mathbb{E}[|V_0^i - \overline{V}_0^i|^2] + \left(\frac{\gamma}{m}\right)^2 \mathbb{E}\left[\left|\int_0^t (V_s^i - \overline{V}_s^i) I_M(s)\, ds\right|^2\right]
$$

$$
+ \left(\frac{\lambda}{m}\right)^2 \mathbb{E}\left[\left|\int_0^t \left((x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i) - (x_\alpha(\rho_{X,s}) - \overline{X}_s^i)\right) I_M(s)\, ds\right|^2\right]
$$

$$
+ \left(\frac{\sigma}{m}\right)^2 \mathbb{E}\left[\left|\int_0^t \left(|x_\alpha(\widehat{\rho}_{X,s}^N) - X_s^i| - |x_\alpha(\rho_{X,s}) - \overline{X}_s^i|\right) I_M(s)\, dB_s^i\right|^2\right]
$$

$$
\leq C\left(\mathbb{E}[|V_0^i - \overline{V}_0^i|^2] + \int_0^t \mathbb{E}\left[|V_s^i - \overline{V}_s^i|^2 I_M(s)\right] ds\right.
$$

$$
\left. + \int_0^t \mathbb{E}\left[\left(|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 + |X_s^i - \overline{X}_s^i|^2\right) I_M(s)\right] ds\right) \qquad (4.12)
$$

with $C = C(\gamma/m, \lambda/m, \sigma/m, T)$. In the first step of (4.12) we used that the processes (X_t^i, V_t^i) and $(\overline{X}_t^i, \overline{V}_t^i)$ share the Brownian motion paths, and in the second both Itô isometry and Jensen's inequality. In order to conclude, it remains to control the term $\mathbb{E}\left[|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 I_M(s)\right]$. To do so, in analogy to the definition of $\widehat{\rho}_{X,s}^N$, let us denote by $\overline{\rho}_{X,s}^N$ the empirical measure associated with the processes $\overline{X}_s^i$, i.e., $\overline{\rho}_{X,s}^N := \frac{1}{N} \sum_{i=1}^N \delta_{\overline{X}_s^i}$. Then, by following the proofs of [8, Lemma 3.2] and [16, Lemma 3.1], and exploiting the boundedness ensured by the multiplication with the random variable $I_M(s)$, we obtain

$$
\mathbb{E}\left[|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 I_M(s)\right]
$$

$$
\lesssim \mathbb{E}\left[|x_\alpha(\widehat{\rho}_{X,s}^N) - x_\alpha(\overline{\rho}_{X,s}^N)|^2 I_M(s)\right] + \mathbb{E}\left[|x_\alpha(\overline{\rho}_{X,s}^N) - x_\alpha(\rho_{X,s})|^2 I_M(s)\right]
$$

$$
\leq C\left(\frac{1}{N} \sum_{i=1}^N \mathbb{E}[|X_s^i - \overline{X}_s^i|^2 I_M(s)] + N^{-1}\right)
$$

$$
\leq C\left(\max_{i=1,\ldots,N} \mathbb{E}[|X_s^i - \overline{X}_s^i|^2 I_M(s)] + N^{-1}\right)
$$

with $C = C(\alpha, L_{\mathcal{E}}, c_{\mathcal{E}}, |x^*|, M, b_1, b_2)$. Inserting the latter into (4.12), and adding up (4.11) and (4.12) yields

$$
\mathbb{E}[(|X_t^i - \overline{X}_t^i|^2 + |V_t^i - \overline{V}_t^i|^2) I_M(t)]
$$

$$
\leq C \int_0^t \mathbb{E}\left[(|X_s^i - \overline{X}_s^i|^2 + |V_s^i - \overline{V}_s^i|^2) I_M(s)\right] \qquad (4.13)
$$

$$
+ \max_{j=1,\ldots,N} \mathbb{E}\left[|X_s^j - \overline{X}_s^j|^2 I_M(s)\right] + N^{-1} ds
$$

with $C = C(\alpha, \gamma/m, \lambda/m, \sigma/m, T, L_{\mathcal{E}}, c_{\mathcal{E}}, |x^*|, M, b_1, b_2)$ and where we used that the processes (X_t^i, V_t^i) and $(\overline{X}_t^i, \overline{V}_t^i)$ share the initial conditions. Lastly, by taking the maximum over i on both sides we get

$$
\max_{i=1,\dots,N} \mathbb{E}\left[\left(|X_t^i - \overline{X}_t^i|^2 + |V_t^i - \overline{V}_t^i|^2 \right) I_M(t) \right]
$$

$$
\leq C \int_0^t \mathbb{E}\left[\max_{j=1,\dots,N} \mathbb{E}\left[\left(|X_s^j - \overline{X}_s^j|^2 + |V_s^j - \overline{V}_s^j|^2 \right) I_M(s) \right] + N^{-1} \right] ds
$$

$$(4.14)$$

with the C from before. After recalling the definition of the conditional expectation, an application of Grönwall's inequality concludes the proof. $\qquad\square$

Remark 8 While the first part of Theorem 5 about the uniform in time boundedness of the empirical measures holds mutatis mutandis for the PSO dynamics with memory effects (1.2) and (1.8), it does not seem straightforward to obtain the second part in this setting due to the way the memory effects are implemented in (1.2b) and (1.8b). As a matter of fact, this is due to exactly the same technical reasons why we lack a uniqueness statement in Sect. 3.1. We therefore leave the investigation of this extension to future research, in particular in regard to the question whether a suitably modified proof technique or another implementations of memory effects resolve this issue.

4.2 Convergence of PSO Without Memory Effects in Probability

Combining Theorem 5 with the convergence analysis of the mean-field dynamics (2.1) as described in Theorem 2, as well as a classical result about the numerical approximation of SDEs allows to obtain convergence guarantees with provable polynomial complexity for the numerical PSO method as stated in Theorem 6 below. Let us, for the reader's convenience, recall from [21, Section 6] that a possible discretized version of the interacting particle system (2.1) is given by

$$
X_{(k+1)\Delta t}^i = X_{k\Delta t}^i + \Delta t \, V_{(k+1)\Delta t}^i, \tag{4.15a}
$$

$$
V_{(k+1)\Delta t}^i = \left(\frac{m}{m + \Delta t \gamma} \right) V_{k\Delta t}^i + \left(\frac{\Delta t \lambda}{m + \Delta t \gamma} \right) \left(x_\alpha(\widehat{\rho}_{X,k\Delta t}^N) - X_{k\Delta t}^i \right)
$$

$$
+ \left(\frac{\sqrt{\Delta t}\sigma}{m + \Delta t \gamma} \right) D\left(x_\alpha(\widehat{\rho}_{X,k\Delta t}^N) - X_{k\Delta t}^i \right) B_{k\Delta t}^i \tag{4.15b}
$$

for $k = 0, \dots, K$ and where $\left((B_{k\Delta t}^i)_{k=1,\dots,K-1} \right)_{i=1,\dots,N}$ are independent, identically distributed standard Gaussian random vectors in $\mathbb{R}^d$.

Theorem 6 *Let $\epsilon_{\text{total}} > 0$ and $\delta \in (0, 1/2)$. Then, under the assumptions of Theorems 2 and 5, it holds for the discretized PSO dynamics (4.15) that*

$$
\left| \frac{1}{N} \sum_{i=1}^N X_{K\Delta t}^i - x^* \right|^2 \leq \epsilon_{\text{total}} \tag{4.16}
$$

with probability larger than $1 - \left(\delta + \epsilon_{\text{total}}^{-1}(C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \widetilde{\varepsilon} + \varepsilon^{2\nu}/\eta^2)\right)$. *Here, m denotes the order of accuracy of the used discretization scheme. Moreover, besides problem dependent factors and the parameters of the method, the dependence of the constants is as follows. C_{NA} depends linearly on d and N, and exponentially on T. C_{MFA} depends on exponentially on α, T and δ^{-1}. C_{LLN} depends on the moment bound from Theorem 1. Lastly, $\widetilde{\varepsilon}$ and ε are chosen according to Theorem 2.*

Remark 9 It is worth emphasizing at this point that the time horizon T in Theorem 6 scales as $\mathcal{O}\left(\log(\widetilde{\varepsilon}^{-1})/\chi\right)$ and therefore logarithmically in the desired accuracy $\widetilde{\varepsilon}$ as a result of Theorem 2, see also the proof below. This ensures that the constants C_{NA} and C_{MFA} appearing implicitly in the bound (4.16) do not lead to an unfeasible numerical method by requiring extremely small time step sizes Δt and an exceedingly large amount of particles N.

Proof of Theorem 6 The overall error can be decomposed as

$$
\mathbb{E}\left[\left|\left|\frac{1}{N}\sum_{i=1}^{N} X_{K\Delta t}^i - x^*\right|\right|^2 \,\bigg|\, \Omega_M\right]
$$
$$
\lesssim \mathbb{E}\left[\left|\left|\frac{1}{N}\sum_{i=1}^{N}\left(X_{K\Delta t}^i - X_T^i\right)\right|\right|^2\right] + \mathbb{E}\left[\left|\left|\frac{1}{N}\sum_{i=1}^{N}\left(X_T^i - \overline{X}_T^i\right)\right|\right|^2 \,\bigg|\, \Omega_M\right] \quad (4.17)
$$
$$
+ \mathbb{E}\left[\left|\left|\frac{1}{N}\sum_{i=1}^{N}\overline{X}_T^i - \mathbb{E}\left[\overline{X}_T\right]\right|\right|^2\right] + \left|\mathbb{E}\left[\overline{X}_T\right] - \widetilde{x}\right|^2 + \left|\widetilde{x} - x^*\right|^2,
$$

where we used that $\mathbb{P}(\Omega_M) \geq (1 - \delta) \geq 1/2$. By means of a classical result about the convergence of numerical schemes for SDEs [43], the first term in (4.17) can be bounded by $C_{\text{NA}}(\Delta t)^m$. For the second term, Theorem 5 gives the estimate $C_{\text{MFA}}N^{-1}$. The third term can be bounded by $C_{\text{LLN}}N^{-1}$ as a consequence of the law of large numbers. Eventually, Theorem 2 allows us to choose $T = \mathcal{O}\left(\log(\widetilde{\varepsilon}^{-1})/\chi\right)$ sufficiently large to reach any prescribed accuracy $\widetilde{\varepsilon}$ for the next-to-last term as well as $\varepsilon^{2\nu}/\eta^2$ for the last term by a suitable choice of α. With these individual bounds we obtain

$$
\mathbb{E}\left[\left|\left|\frac{1}{N}\sum_{i=1}^{N} X_{K\Delta t}^i - x^*\right|\right|^2 \,\bigg|\, \Omega_M\right]
$$
$$
\leq C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \widetilde{\varepsilon} + \varepsilon^{2\nu}/\eta^2. \quad (4.18)
$$

It now remains to estimate the probability of the set $K_{\epsilon_{\text{total}}}^N \subset \Omega$, where Inequality (4.16) does not hold. By utilizing the conditional version of Markov's inequality together with

the formerly established bound (4.18), we have

$$
\begin{aligned}
\mathbb{P}\big(K_{\epsilon_{\text{total}}}^N\big) &= \mathbb{P}\big(K_{\epsilon_{\text{total}}}^N \cap \Omega_M\big) + \mathbb{P}\big(K_{\epsilon_{\text{total}}}^N \cap \Omega_M^c\big) \\
&\le \mathbb{P}\big(K_{\epsilon_{\text{total}}}^N \,\big|\, \Omega_M\big)\,\mathbb{P}(\Omega_M) + \mathbb{P}\big(\Omega_M^c\big) \\
&\le \frac{C_{\text{NA}}(\Delta t)^m + C_{\text{MFA}}N^{-1} + C_{\text{LLN}}N^{-1} + \widetilde{\varepsilon} + \varepsilon^{2\nu}/\eta^2}{\epsilon_{\text{total}}} + \delta
\end{aligned}
\tag{4.19}
$$

for a sufficiently large choice of M in (4.1).   □

A result in this spirit was first presented for CBO in [18, Theorem 14] and is hereby extended to PSO.

5 Implementation of PSO and Numerical Results

The purpose of this section is twofold. At first, an efficient implementation of PSO is provided, which is particularly suited for high-dimensional optimization problems arising in machine learning. Its performance is then evaluated on a standard benchmark problem, where we investigate the influence of the parameters, and the training of a neural network classifier for handwritten digits. Furthermore, several relevant implementational aspects are discussed, including the computational complexity and scalability, modifications inspired from simulated annealing and evolutionary algorithms, and the numerical stability of the method.

5.1 An Efficient Implementation of PSO

Let us stress that PSO is an extremely versatile, flexible and customizable optimization method, which can be regarded as a black-box optimizer. As a zero-order method it is not reliant on the gradient information and can be even applied to discontinuous objectives, making it inevitably superior to first-order optimization methods in cases where derivatives are computationally infeasible. However, also in machine learning applications, where gradient-based optimizers are considered the state of the art, PSO may be of particular interest in view of vanishing or exploding gradient phenomena.

Typical objective functions appearing in machine learning are of the form

$$
\mathcal{E}(x) = \frac{1}{M} \sum_{j=1}^{M} \mathcal{E}_j(x),
\tag{5.1}
$$

where $\mathcal{E}_j$ is usually the loss of the jth training sample. In order to run the scheme (1.2), frequent evaluations of $\mathcal{E}$ are necessary, which may be computationally intense or even prohibitive in some applications.

Computational complexity: Inspired by mini-batch gradient descent, the authors of [28] developed a random batch method for interacting particle systems, which was employed for CBO in [9]. In the same spirit, we present with Algorithm 1 a

---

Algorithm 1 Particle swarm optimization (PSO)

Input: Objective $\mathcal{E}$ as in (5.1), time horizon T or number of epochs #*epochs*, discrete time step size Δt, batch sizes n_N and $n_{\mathcal{E}}$, parameters $m, \gamma, \lambda_1, \lambda_2, \sigma_1, \sigma_2, \alpha, \beta, \theta$ and κ, number of particles N, initialization f_0

Output: Approximation $y_\alpha(\widehat{\rho}^N_{Y,T})$ of the global minimizer x^* of $\mathcal{E}$

1: Generate the particles' initial positions and velocities $(X_0^i, V_0^i)_{i=1,\ldots,N}$ according to a common initial law f_0. Initialize the local best positions $Y_0^i = X_0^i$.
2: Ensure that $n_{\mathcal{E}}$ divides M and n_N divides N.
3: Convert T into #*epochs* or vice versa via $T = \#epochs\,(M/n_{\mathcal{E}})(N/n_N)\Delta t$. Set $k = 0$ and $epoch = 1$.
4: **while** $epoch \leq$ #*epochs* and stopping criterion not fulfilled
5:     Partition $\{1, \ldots, M\}$ into batches $\mathcal{B}_k^1, \ldots, \mathcal{B}_k^{M/n_{\mathcal{E}}}$ of batch size $n_{\mathcal{E}}$.
6:     **for** $b = 1, \ldots, M/n_{\mathcal{E}}$
7:         Define the objective function on this batch as

$$\mathcal{E}_{batch}(x) = \frac{1}{n_{\mathcal{E}}} \sum_{j \in \mathcal{B}_k^b} \mathcal{E}_j(x). \tag{5.2}$$

8:         Partition the particles, i.e., the set $\{1, \ldots, N\}$, into disjoint sets $\mathcal{P}_k^1, \ldots, \mathcal{P}_k^{N/n_N}$ of size n_N.
9:         **for** $n = 1, \ldots, N/n_N$
10:             Compute the consensus point $y_\alpha(\widehat{\rho}^{N/n_N}_{Y,k\Delta t})$ according to Equation (1.4) with objective $\mathcal{E}_{batch}$ from the particles in $\mathcal{P}_k^n$, i.e., with the empirical measure $\widehat{\rho}^{N/n_N}_{Y,k\Delta t} = \frac{1}{n_N} \sum_{i \in \mathcal{P}_k^n} \delta_{Y_{k\Delta t}^i}$.
11:             Update either all particles (full update) or only the particles in the current batch $\mathcal{P}_k^n$ (partial update) according to a discretized version of the PSO dynamics (1.2).
12:             **if** $k > 0$ and $\left| y_\alpha(\rho^N_{Y,k\Delta t}) - y_\alpha(\rho^N_{Y,(k-1)\Delta t}) \right|$ is too small despite stopping criterion not fulfilled
13:                 Perform an independent Brownian motion for the positions or velocities of all particles.
14:             **end if**
15:             Set $k = k + 1$.
16:         **end for**
17:     **end for**
18:     Check the stopping criterion and **break** if fulfilled. If not, employ the optional strategies described at the end of Section 5.1, set $epoch = epoch + 1$ and continue.
19: **end while**
20: Compute the consensus point $y_\alpha(\widehat{\rho}^N_{Y,T})$ according to Equation (1.4) with objective $\mathcal{E}$ from all particles, i.e., with $\widehat{\rho}^N_{Y,T} = \frac{1}{N} \sum_{i=1}^N \delta_{Y_T^i}$.

---

computationally efficient implementation of PSO. The mini-batch idea is present on two different levels. In line 7, the objective is defined with respect to a batch of the training data of size $n_{\mathcal{E}}$, meaning that only a subsample of the data is considered. One epoch is completed after each data sample was seen exactly once, i.e., after $M/n_{\mathcal{E}}$ steps. At each step the consensus point y_α has to be computed, for which $\mathcal{E}_{batch}$ needs to be evaluated for N particles. This still constitutes the most significant computational effort. However, the mini-batch idea can be leveraged for a second time. In the **for** loop in line 9 we partition the particles into sets of size n_N and perform the updates of line 11 only for the n_N particles in the respective subset. Since this is embarrassingly parallel, a parallel machine can be deployed to reduce the runtime by up to a factor p (the number of available processors). While this is referred to as partial update, alternatively, on a sequential architecture, a full update can be made at every iteration, requiring all N particles to be updated in line 11. Apart from lowering the required computing

resources tremendously, these mini-batch ideas actually improve the stability of the method and the capability of finding good optima by introducing more stochasticity into the algorithm.

Concerning additional computational complexity due to the usage of memory effects, let us point out that, except for the required storage of the local (historical) best positions and their objective values, the update rule (1.3) in combination with the partial update allows to include such mechanisms with no additional cost by keeping track of the objective values of the local best positions. In such case, only one function call of each $\mathcal{E}_{batch}$ per epoch and per particle is necessary, which is optimal and coincides with PSO without memory effects or CBO. A different realization of (1.2b) might result in a higher cost.

Implementational aspects: A discretization of the SDE (1.2) in line 11 can be obtained for instance from a simple Euler-Maruyama or semi-implicit scheme [23, 43], see, e.g., [21, Equation (6.3)]. In our numerical experiments below Equation (1.3) is used for updating the local best position, which corresponds to $\kappa = 1/(2\Delta t)$, $\theta = 0$, and $\beta = \infty$. Furthermore, the friction parameter is set according to $\gamma = 1 - m$, which is a typical choice in the literature. Let us also remark that a numerically stable computation of the consensus point in lines 10 and 20 for $\alpha \gg 1$ can be obtained by replacing $\mathcal{E}_{batch}$ with $\mathcal{E}_{batch} - \widetilde{\underline{\mathcal{E}}}$, where $\widetilde{\underline{\mathcal{E}}} := \min_{i \in \mathcal{P}_k^n} \mathcal{E}_{batch}(Y_{k\Delta t}^i)$.

Cooling and evolutionary strategies: The PSO algorithm can be divided into two phases, an exploration phase, where the energy landscape is searched coarsely, and a determination phase, where the final output is identified. While the former benefits from small α and large diffusion parameters, in the latter, $\alpha \gg 1$ guarantees the selection of the best solution. A cooling strategy inspired from simulated annealing allows to start with moderate α and relatively large diffusion parameters σ_1, σ_2. After each epoch, α is multiplied by 2, while the diffusion parameters follow the schedule $\sigma = \sigma/\log(epoch + 2)$ for $\sigma \in \{\sigma_1, \sigma_2\}$. Such strategy was proposed in [9, Section 4] for CBO. In order to further reduce computational complexity, the provable decay of the variance suggests to decrease the number of agents by discarding particles in accordance with the empirical variance. A possible schedule for the number of agents proposed in [20, Section 2.2] is to set $N_{epoch+1} = \lceil N_{epoch}((1 - \mu) + \mu \widetilde{\Sigma}_{epoch}/\Sigma_{epoch}) \rceil$ for $\mu \in [0, 1]$ and where Σ_{epoch} and $\widetilde{\Sigma}_{epoch}$ denote the empirical variances of the N_{epoch} particles at the beginning and at the end of the current epoch.

5.2 Numerical Experiments for the Rastrigin Function

Before turning to high-dimensional optimization problems, let us discuss the parameter choices of PSO in moderate dimensions ($d = 20$) at the example of the well-known Rastrigin benchmark function $\mathcal{E}(v) = \sum_{k=1}^{d} v_k^2 + \frac{5}{2}(1 - \cos(2\pi v_k))$, which meets the requirements of Assumption 1 despite being highly non-convex with many spurious local optima. To narrow down the number of tunable parameters, we let $\gamma = 1 - m$, choose $\alpha = 100$, $N = 100$, and update the local best position (if present) according to Equation (1.3), i.e., $\kappa = 1/(2\Delta t)$, $\theta = 0$, and $\beta = \infty$. We moreover let $\lambda_2 =$
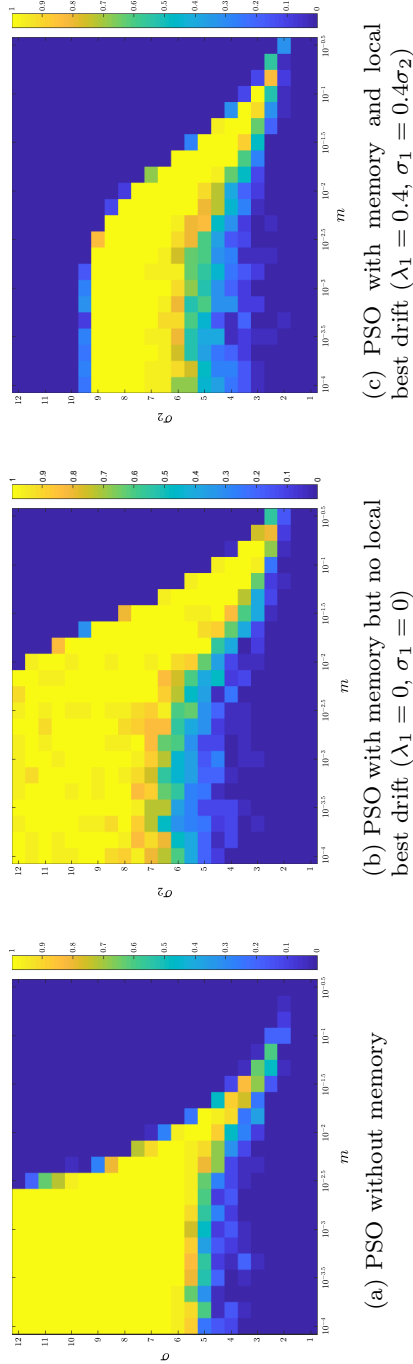
(a) PSO without memory

(b) PSO with memory but no local best drift ($\lambda_1 = 0$, $\sigma_1 = 0$)

(c) PSO with memory and local best drift ($\lambda_1 = 0.4$, $\sigma_1 = 0.4\sigma_2$)

Fig. 2 Phase transition diagrams comparing PSO without and with memory effects for different inertia parameters m and noise coefficients σ (PSO without memory) and σ_2 (PSO with memory). The empirical success probability is computed from 25 runs and depicted by color from zero (blue) to one (yellow)

(a) Dense shallow NN.

(b) CNN with two convolutional and two pooling layers, and one dense layer.
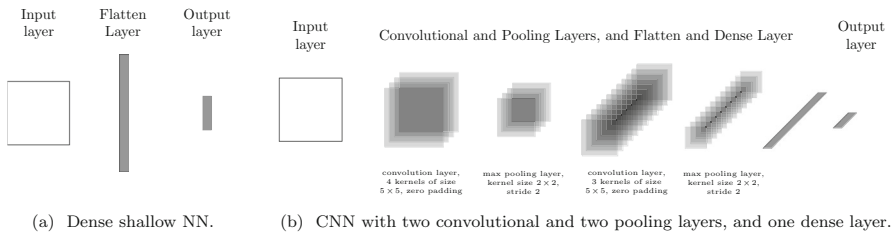
Fig. 3 Architectures of the NNs used in the experiments of Sect. 5.3, cf. [19, Section 4]

1 (or $\lambda = 1$ for PSO without memory) and $\Delta t = 0.01$, which are such that the algorithm either finds consensus or explodes within the time horizon $T = 100$ in all instances. For simplicity we assume that $\sigma_1 = \lambda_1 \sigma_2$. The algorithm is initialized with positions distributed according to $\mathcal{N}\big((2, \ldots, 2), 4\mathrm{Id}\big)$ and velocities according to $\mathcal{N}\big((0, \ldots, 0), \mathrm{Id}\big)$. In Fig. 2 we depict the phase diagram comparing the success probability of PSO for different parameter choices of the inertia parameter m and the diffusion parameter σ or σ_2, respectively.

We observe that for m fixed there is a noise threshold above which the dynamics explodes. While smaller m permit a larger flexibility in the used noise, they require an individual minimal noise level. Further numerical experiments suggest however that increasing the number of particles N allows for a lower minimal noise level. There are subtle differences between PSO without and with memory, but they are not decisive as in applications also confirmed by the numerical experiments in Sect. 5.3, [22, Section 5.3] as well as the survey paper [21, Section 6.3].

5.3 A Machine Learning Application

We now showcase the practicability of PSO as implemented in Algorithm 1 at the example of a very competitive high-dimensional benchmark problem in machine learning, the classification of handwritten digits. In what follows we train a shallow and a convolutional NN (CNN) classifier for the MNIST dataset [34]. Let us point out, that it is not our objective to challenge the state of the art by employing the most sophisticated model (deep CNNs achieve near-human performance of more than 99.5% accuracy). Instead, we want to demonstrate that PSO reaches results comparable to SGD with backpropagation, while at the same time relying exclusively on the evaluation of $\mathcal{E}$.

In our experiment we use NNs with architectures as depicted in Fig. 3.

The input is a black-and-white image represented by a (28×28)-dimensional matrix with entries between 0 and 1. For the shallow NN (see Fig. 3a), the flattened image is passed through a dense layer $\mathrm{ReLU}(W \cdot + b)$ with trainable weights $W \in \mathbb{R}^{10 \times 728}$ and bias $b \in \mathbb{R}^{10}$. Our CNN (see Fig. 3b) is similar to LeNet-1, cf. [33, Section III.C.7]. Each dense or convolution layer has a ReLU activation and is followed by a batch normalization layer to speed up the training process. Eventually, the final layers of both NNs apply a softmax activation function allowing to interpret the 10-dimensional output vector as a probability distribution over the digits.
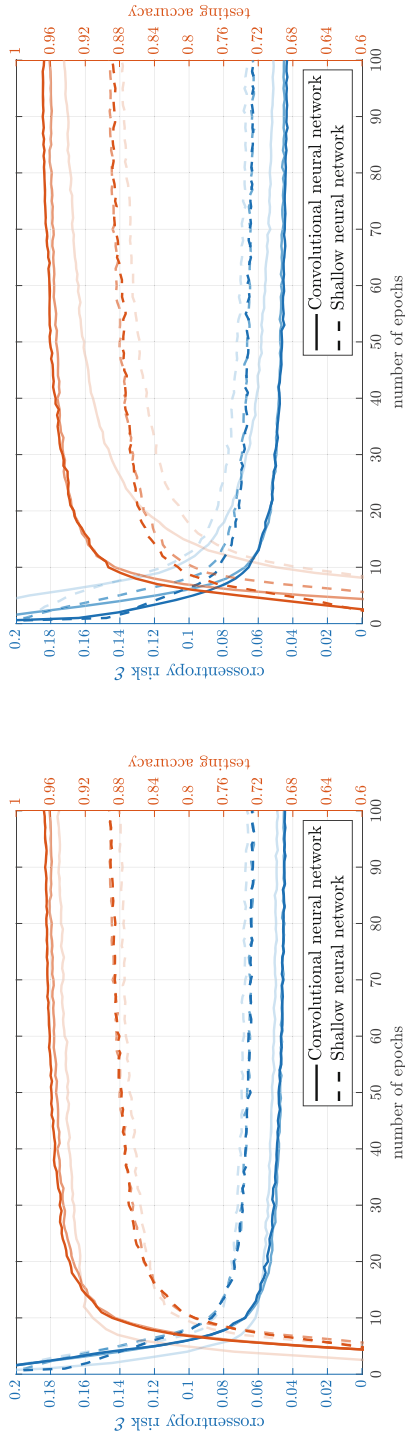
(a) PSO for three different memory settings: without memory (lightest lines); with memory but without local best drift, i.e., $\lambda_1 = 0$, $\sigma_1 = 0$ (line with intermediate opacity); with memory with local best drift $\lambda_1 = 0.4$, $\sigma_1 = \lambda_1 \sigma_2$ (darkest lines).

(b) PSO with memory but without local best drift for three different inertia parameters $m \in \{0.1, 0.2, 0.4\}$ (increasing opacity for larger m).
Note that, for reference, the lines with intermediate opacity coincide with the ones of Fig. 4a.

Fig. 4 Comparison of the performances of a shallow (dashed lines) and convolutional (solid lines) NN with architectures as described in Fig. 3, when trained with PSO as in Algorithm 1. Depicted are the accuracies on a test dataset (orange lines) and the values of the objective function $\mathcal{E}$ (blue lines) evaluated on a random sample of the training set of size 10000

We denote by θ the trainable parameters of the NNs, which are 7850 for the shallow NN and 2112 for the CNN. They are learned by minimizing $\mathcal{E}(\theta) = \frac{1}{M} \sum_{j=1}^{M} \ell(f_\theta(x^j), y^j)$, where f_θ denotes the forward pass of the NN, (x^j, y^j) the jth image-label tuple and ℓ the categorical crossentropy loss $\ell(\widehat{y}, y) = -\sum_{k=0}^{9} y_k \log(\widehat{y}_k)$. The performance is measured by counting the number of successful predictions on a test set. We use a train-test split of 60000 training and 10000 test images. For our experiments we choose $\lambda_2 = 1$, $(\sigma_2)_{initial} = \sqrt{0.4}$, $\alpha_{initial} = 50$, $\Delta t = 0.1$ and update the local best position according to Equation (1.3). We use $N = 100$ agents, which are initialized according to $\mathcal{N}\big((0, \ldots, 0)^T, \mathrm{Id}\big)$ in position and velocity. The mini-batch sizes are $n_\mathcal{E} = 60$ and $n_N = 100$ (consequently a full update is performed in line 11) and a cooling strategy is used in line 18.

Figure 4a reports the performances for different memory settings and fixed $m = 0.2$, whereas Fig. 4b depicts the results for different inertia parameters m in the case of PSO with memory but no memory drift.

For the shallow NN, we obtain a test accuracy of above 89%, whereas the CNN achieves almost 97%. To put those numbers into perspective, when trained with SGD, a similar performance for the shallow NN, see [9, Figure 7], and a benchmark accuracy of 98.3% for a comparable CNN, cf. [33, Figure 9], are reached. As can be seen from Fig. 4a, the usage of the local best positions when computing the consensus point significantly improves the performance. The advantage of having an additional drift towards the local best position is less pronounced. Regarding the inertia parameter m in Fig. 4b, our numerical results suggest that larger m yield faster convergence.

6 Conclusions

In this paper we prove the convergence of PSO without and with memory effects to a global minimizer of a possibly nonconvex and nonsmooth objective function in the mean-field sense. Our analysis holds under a suitable well-preparation condition about the initialization and comprises a rich class of objectives which in particular includes functions with multiple global minimizers. For PSO without memory effects we furthermore quantify how well the mean-field dynamics approximates the interacting finite particle dynamics, which is implemented for numerical experiments. Since in particular the latter results does not suffer from the curse of dimensionality, we thereby prove that the numerical PSO method has polynomial complexity. With this we contribute to the completion of a mathematically rigorous understanding of PSO. Furthermore, we propose a computationally efficient and parallelizable implementation and showcase its practicability by training a CNN reaching a performance comparable to stochastic gradient descent.

It remains an open problem to extend the mean-field approximation result to the variant of PSO with memory effects or, alternatively, to devise an implementation of such effects compatible with the used proof technique. Moreover, we also leave a more thorough understanding of the influence of the parameters as well as the influence of memory effects for future, more experimental research.

Finally, we believe that the analysis framework of this and prior works on CBO [8, 18, 42] motivates to investigate also other renowned metaheuristic algorithms through the lens of a mean-field limit.

Declarations

Competing interest The authors declare that they have no conflict of interest.

References

1. Aarts, E., Korst, J.: Simulated annealing and Boltzmann machines. A stochastic approach to combinatorial optimization and neural computing. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley , Chichester (1989)
2. Arnold, L.: Stochastic differential equations: Theory and applications. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney (1974). Translated from the German
3. Azencott, R. (ed.): Simulated annealing: Parallelization techniques. Wiley (1992)
4. van den Bergh, F.: An analysis of particle swarm optimizers. Ph.D. thesis, University of Pretoria (2007)
5. van den Bergh, F., Engelbrecht, A.P.: A convergence proof for the particle swarm optimiser. Fund. Inform. **105**(4), 341–374 (2010)
6. Bolley, F., Cañizo, J.A., Carrillo, J.A.: Stochastic mean-field limit: non-Lipschitz forces and swarming. Math. Models Methods Appl. Sci. **21**(11), 2179–2210 (2011)
7. Bruned, V., Mas, A., Wlodarczyk, S.: Weak convergence of particle swarm optimization. arXiv preprint arXiv:1811.04924 (2018)
8. Carrillo, J.A., Choi, Y.P., Totzeck, C., Tse, O.: An analytical framework for consensus-based global optimization method. Math. Models Methods Appl. Sci. **28**(6), 1037–1066 (2018)
9. Carrillo, J.A., Jin, S., Li, L., Zhu, Y.: A consensus-based global optimization method for high dimensional machine learning problems. ESAIM Control Optim. Calc. Var. **27**(suppl.), Paper No. S5, 1–22 (2021)
10. Cipriani, C., Huang, H., Qiu, J.: Zero-inertia limit: from particle swarm optimization to consensus-based optimization. SIAM J. Math. Anal. **54**(3), 3091–3121 (2022)
11. Clerc, M., Kennedy, J.: The particle swarm—explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evol. Comput. **6**(1), 58–73 (2002). https://doi.org/10.1109/4235.985692
12. Dembo, A., Zeitouni, O.: Large deviations techniques and applications, Applications of Mathematics (New York), vol. 38, 2nd edn. Springer, New York (1998)
13. Ding, Z., Chen, S., Li, Q., Wright, S.: On the global convergence of gradient descent for multi-layer resnets in the mean-field regime. arXiv preprint arXiv:2110.02926 (2021)

14. Dréo, J., Pétrowski, A., Siarry, P., Taillard, E.: Metaheuristics for hard optimization. Springer, Berlin (2006). Methods and case studies, Translated from the 2003 French original by Amitava Chatterjee
15. Fogel, D.B.: Evolutionary Computation. Toward a New Philosophy of Machine Intelligence, 2nd edn. IEEE Press, Piscataway (2000)
16. Fornasier, M., Huang, H., Pareschi, L., Sünnen, P.: Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. Math. Models Methods Appl. Sci. **30**(14), 2725–2751 (2020)
17. Fornasier, M., Huang, H., Pareschi, L., Sünnen, P.: Anisotropic diffusion in consensus-based optimization on the sphere. SIAM J. Optim. **32**(3), 1984–2012 (2022)
18. Fornasier, M., Klock, T., Riedl, K.: Consensus-based optimization methods converge globally. arXiv preprint arXiv:2103.15130 (2021)
19. Fornasier, M., Klock, T., Riedl, K.: Convergence of anisotropic consensus-based optimization in mean-field law. In: J.L. Jiménez Laredo, J.I. Hidalgo, K.O. Babaagba (eds.) Applications of Evolutionary Computation, pp. 738–754. Springer, Cham (2022)
20. Fornasier, M., Pareschi, L., Huang, H., Sünnen, P.: Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. J. Mach. Learn. Res. **22**(237), 1–55 (2021)
21. Grassi, S., Huang, H., Pareschi, L., Qiu, J.: Mean-field particle swarm optimization. arXiv preprint arXiv:2108.00393 (2021)
22. Grassi, S., Pareschi, L.: From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. Math. Models Methods Appl. Sci. **31**(8), 1625–1657 (2021)
23. Higham, D.J.: An algorithmic introduction to numerical simulation of stochastic differential equations. SIAM Rev. **43**(3), 525–546 (2001)
24. Holland, J.H.: Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, Ann Arbor, Mich. (1975)
25. Huang, H., Liu, J.G., Pickl, P.: On the mean-field limit for the Vlasov-Poisson-Fokker-Planck system. J. Stat. Phys. **181**(5), 1915–1965 (2020)
26. Huang, H., Qiu, J.: On the mean-field limit for the consensus-based optimization. Mathematical Methods in the Applied Sciences, pp. 1–18 (2022)
27. Jabin, P.E., Wang, Z.: Mean field limit for stochastic particle systems. In: Active particles. Vol. 1. Advances in theory, models, and applications, Model. Simul. Sci. Eng. Technol., pp. 379–402. Birkhäuser/Springer, Cham (2017)
28. Jin, S., Li, L., Liu, J.G.: Random batch methods (RBM) for interacting particle systems. J. Comput. Phys. **400**(108877), 1–30 (2020)
29. Kadanoff, L.P.: More is the same; phase transitions and mean field theories. J. Stat. Phys. **137**(5–6), 777–797 (2009)
30. Kennedy, J.: The particle swarm: social adaptation of knowledge. In: Proceedings of 1997 IEEE International Conference on Evolutionary Computation, pp. 303–308. IEEE (1997). 10.1109/ICEC.1997.592326
31. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE (1995). 10.1109/ICNN.1995.488968
32. Kushner, H., Yin, G.G.: Stochastic approximation and recursive algorithms and applications, vol. 35. Springer Science & Business Media (2003)
33. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
34. LeCun, Y., Cortes, C., Burges, C.: MNIST handwritten digit database (2010). http://yann.lecun.com/exdb/mnist/
35. Lin, S.W., Ying, K.C., Chen, S.C., Lee, Z.J.: Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Syst. Appl. **35**(4), 1817–1824 (2008)
36. Mei, S., Montanari, A., Nguyen, P.M.: A mean field view of the landscape of two-layer neural networks. Proc. Natl. Acad. Sci. U.S.A. **115**(33), E7665–E7671 (2018)
37. Miclo, L.: Recuit simulé sur $\mathbb{R}^n$. étude de l'évolution de l'énergie libre. In: Annales de l'IHP Probabilités et statistiques, vol. 28, pp. 235–266 (1992)
38. Miller, P.D.: Applied Asymptotic Analysis, Graduate Studies in Mathematics, vol. 75. American Mathematical Society, Providence, RI (2006)
39. Øksendal, B.: Stochastic Differential Equations: An Introduction with Applications, 6th edn. Springer, Berlin (2003)
40. Özcan, E., Mohan, C.K.: Analysis of a simple particle swarm optimization system (1998)

41. Panigrahi, B.K., Shi, Y., Lim, M.H.: Handbook of swarm intelligence: concepts, principles and applications, vol. 8. Springer Science & Business Media (2011)
42. Pinnau, R., Totzeck, C., Tse, O., Martin, S.: A consensus-based model for global optimization and its mean-field limit. Math. Models Methods Appl. Sci. **27**(1), 183–204 (2017)
43. Platen, E.: An introduction to numerical methods for stochastic differential equations. In: Acta numerica, Acta Numer., vol. 8, pp. 197–246. Cambridge University Press, Cambridge (1999)
44. Poli, R.: Mean and variance of the sampling distribution of particle swarm optimizers during stagnation. IEEE Trans. Evol. Comput. **13**(4), 712–721 (2009)
45. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intell **1**(1), 33–57 (2007). https://doi.org/10.1007/s11721-007-0002-0
46. Rastrigin, L.A.: The convergence of the random search method in the external control of many-parameter system. Autom. Remote Control **24**, 1337–1342 (1963)
47. Revuz, D., Yor, M.: Continuous martingales and Brownian motion, Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, third edn. Springer-Verlag, Berlin (1999)
48. Riedl, K.: Leveraging memory effects and gradient information in consensus-based optimization: On global convergence in mean-field law. arXiv:2211.12184 (2022)
49. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
50. Royer, G.: An initiation to logarithmic Sobolev inequalities. 5. American Mathematical Soc. (2007)
51. Schmitt, M., Wanka, R.: Particle swarm optimization almost surely finds local optima. Theor. Comput. Sci. **561**(Part A), 57–72 (2015)
52. Sznitman, A.S.: Topics in propagation of chaos. In: École d'Été de Probabilités de Saint-Flour XIX – 1989, Lecture Notes in Math., vol. 1464, pp. 165–251. Springer, Berlin (1991)
53. Tang, W., Zhou, X.Y.: Tail probability estimates of continuous-time simulated annealing processes. Numerical Algebra, Control and Optimization (2022)
54. Totzeck, C., Wolfram, M.T.: Consensus-based global optimization with personal best. Math. Biosci. Eng. **17**(5), 6026–6044 (2020)
55. Witt, C.: Theory of particle swarm optimization. In: Theory of randomized search heuristics, Ser. Theor. Comput. Sci., vol. 1, pp. 197–223. World Sci. Publ., Hackensack, NJ (2011)
56. Yuan, Q., Yin, G.: Analyzing convergence and rates of convergence of particle swarm optimization algorithms using stochastic approximation methods. IEEE Trans. Autom. Control **60**(7), 1760–1773 (2014)
57. Zhang, Y., Wang, S., Ji, G.: A comprehensive survey on particle swarm optimization algorithm and its applications. Math. Probl. Eng. **931256**, 1–38 (2015)