MEAN-FIELD PARTICLE SWARM OPTIMIZATION

Sara Grassi

Department of Mathematical, Physical and Computer Sciences, University of Parma, Parco Area delle Scienze 7/A, 43124 Parma, Italy sara.grassi@unipr.it

Hui Huang

Department of Mathematics and Statistics, University of Calgary, 2500 University Dr NW, Calgary, AB T2N 1N4, Canada hui.huang1@ucalgary.ca

Lorenzo Pareschi

Department of Mathematics and Computer Science, University of Ferrara, Via Machiavelli 30, 44121 Ferrara, Italy lorenzo.pareschi@unife.it

Jinniao Qiu

Department of Mathematics and Statistics, University of Calgary, 2500 University Dr NW, Calgary, AB T2N 1N4, Canada jinniao.qiu@ucalgary.ca

In this work we survey some recent results on the global minimization of a non-convex and possibly non-smooth high dimensional objective function by means of particle based gradient-free methods. Such problems arise in many situations of contemporary interest in machine learning and signal processing. After a brief overview of metaheuristic methods based on particle swarm optimization (PSO), we introduce a continuous formulation via second-order systems of stochastic differential equations that generalize PSO methods and provide the basis for their theoretical analysis. Subsequently, we will show how through the use of mean-field techniques it is possible to derive in the limit of large particles number the corresponding mean-field PSO description based on Vlasov-Fokker-Planck type equations. Finally, in the zero inertia limit, we will analyze the corresponding macroscopic hydrodynamic equations, showing that they generalize the recently introduced consensus-based optimization (CBO) methods by including memory effects. Rigorous results concerning the mean-field limit, the zero-inertia limit, and the convergence of the mean-field PSO method towards the global minimum are provided along with a suite of numerical examples.

Contents

1	Introduction		
2	Second order stochastic models for particle swarm optimization		6
	2.1	The standard PSO algorithm	7
	2.2	The stochastic differential PSO system	8
	2.3	Stochastic differential PSO model with inertia	9
3	Mean-field particle swarm optimization		10
	3.1	Regularized PSO dynamics without memory effects	11
		3.1.1 Mean-field limit	13
	3.2	Regularized PSO dynamic with memory and local best	21
4	Zero-inertia limit and consensus-based optimization		22
	4.1	The case without memory effects	22
		4.1.1 Formal derivation in the mean-field case	23
		4.1.2 Rigorous derivation	25
	4.2	The general case with memory	37
		4.2.1 Formal derivation in the mean-field case	38
		4.2.2 Rigorous derivation	39
5	Con	vergence to the global minimum	40
6	Numerical examples		47
	6.1	Validation of the mean field limit	48
		6.1.1 Absence of memory effects	49
		6.1.2 Only local best dynamics	50
		6.1.3 The general case	51
	6.2	Numerical small inertia limit	52
	6.3	Performance on high-dimensional test cases	53
		6.3.1 Effect of the inertial parameter m	54
		6.3.2 Effect of the local best dynamics	56
		6.3.3 Comparison on prototype functions	57
7	Con	cluding remarks and research directions	59
References			62

1. Introduction

The Particle Swarm Optimization (PSO) algorithm was introduced by James Kennedy, a social psychologist, and Russel Eberhart, an electrical engineer, in the mid-1990s [1,2]. Since its introduction, due to its simplicity and versatility, the PSO method has gained a great deal of attention from the scientific community, resulting in a huge number of variants of the original algorithm [3–7]. The origin of the method can actually be traced back to an earlier time, since the basic principle of optimization by interacting agents is inspired by previous attempts to reproduce the observed behaviors of animals in their natural habitat, such as flocks of birds or schools of fish [8–13]. These roots in the natural processes of collective animal behavior lead to the PSO algorithm's classification as belonging to Swarm Intelligence (SI), where the notion of swarm intelligence refers to the property of a system in which the coordinated behaviors of agents interacting locally with their environment cause coherent global functional patterns (e.g., self-organization, emergent behavior) to emerge [14–18].

Currently, similar to other gradient-free approaches [19–25], PSO is considered an efficient metaheuristic method for solving complex optimization problems and is available in several programming language libraries. Gradient-based optimizers are effective at finding local minima for high-dimensional, nonlinearly constrained convex problems; however, most gradient-based optimizers have problems dealing with noisy, discontinuous functions, and are not designed to handle discrete and mixed discretecontinuous variables. Unlike gradient-based methods in a convex search space, metaheuristic methods are not necessarily guaranteed to find true global optimal solutions, but they are capable of finding many good solutions that are sometimes sufficient in practical applications. Some of the most popular stochastic metaheuristic methods include Simulated Annealing (SA) [26–28], Ant Colony Optimization (ACO) [29,30], Genetic Algorithms (GA) [31,32] and Differential Evolution (DE) [33,34]. See also [22] for a recent survey on other natured inspired metaheuristics. It should also be mentioned that a large number of newer metaheuristic methods have begun to attract criticism in the research community for hiding their lack of novelty behind elaborate constructions unsupported by any theoretical analysis [35].

In spite of its apparent simplicity, PSO poses formidable challenges for those interested in understanding swarm intelligence through theoretical analysis. To date a fully complete mathematical theory for particle swarm optimization is still lacking (see for example [36–41] and the references therein). The algorithm explores the search space in an intelligent way thanks to a population of particles interacting with each other and updated at each step their position and velocity. Thus, from the theoretical point of view, one can take advantage of the fact that PSO is inspired by classical second order Newtonian dynamics of particle systems. This allows approaches derived from statistical mechanics and mean-field theory to be adapted to the study of the system properties in the limit of a large number of particles [42–49].

Analogies with mean-field dynamics in consensus formation have recently inspired Consensus-based Optimization (CBO) methods, a novel class of particle based methods for global optimization (see [50–58] and the recent survey [59]). Global optimization methods with similar features, but based on Kuramoto-Vicseck dynamics constrained to hypersurfaces [60–62] or on binary Boltzmann dynamics [63], have been introduced and studied recently. These methods are inherently simpler than PSO methods since they were inspired by first order consensus-like dynamics typical of social interactions such as opinion formations and wealth exchanges [64, 65]. In contrast to classic metaheuristic methods typically formulated through a discrete sequence of operations and for which it is quite difficult to provide rigorous convergence to global minimizers, CBO-like methods, thanks to their formulation through stochastic differential equations (SDE) permit to exploit mean-field techniques to prove global convergence for a large class of optimization problems [51,52,61,62]. On the other hand, CBO methods seem to be powerful and robust enough to tackle many interesting high dimensional non-convex optimization problems of interest in machine learning and sampling [52, 57, 61–63, 66, 67].

In this work we review some recent results on the mean-field modeling of particle swarm optimization with the goal of providing a robust mathematical theory for PSO methods and their convergence to the global minimum, based on a continuous description of their dynamics [67–72]. A major difficulty in the mathematical description of PSO methods, and other metaheuristic algorithms, is the presence of memory mechanisms that make their interpretation in terms of differential equations particularly challenging. To this end, the discrete PSO method is generalized via a system of second-order SDEs in which an additional state variable takes into account the memory of the individual particle. We refer to [53] for alternative approaches to memory mechanisms in CBO system.

Adopting the same regularization process for the global best as in CBO

methods [50, 51], it is then possible to pass to the mean-field limit and derive the corresponding Vlasov-Fokker-Planck equation that characterizes the behavior of the system in the limit of a large number of particles [68, 70]. The new mathematical formalism based on mean-field equations permits to study the behavior of the Vlasov-Fokker-Planck PSO model in the limit of zero inertia (see [73–78] for related results in other contexts). In particular, we prove that in this limit the PSO dynamics is described by simplified macroscopic models that correspond to a generalization of CBO models including memory effects and local best [68, 71]. The convergence of the mean-field PSO model to the global minimum is then discussed and shown rigorously in absence of memory effects [72].

Several numerical examples are reported to validate of the mean-field process and the small inertia limit, and to illustrate the role of the various parameters involved in solving high dimensional global optimization problems for various prototype test functions. Other than the basic algorithmic aspects of implementing these generalized PSO methods, we do not discuss the practical algorithmic improvements that can be adopted to increase the success rate, like for example the use of random batch methods [52,79,80], particle reduction techniques [61,62] and parameters adaptivity [4,7]. We refer to [67] for further details on these implementation aspects.

The rest of the survey is organized as follows. In Section 2 we introduce the PSO algorithms and derive the corresponding representations as SDEs using a time continuous approximation of the memory process. Next, in Section 3, thanks to a regularization of the global best and the local best we discuss the large particle limit and derive the respective Vlasov-Fokker-Planck equations describing the mean-field dynamic. A rigorous proof of the mean-field limit is also given. Section 4 is then dedicated to the zeroinertia limit for the mean-field system that allows to recover a CBO model with local best as the corresponding macroscopic limit. This is shown rigorously at the end of the Section. A general convergence result to the global minimum is illustrated in Section 5 in absence of memory effects. Several numerical examples, validating the mean-field approximation, the small inertia limit and testing the performances of the minimizers against some prototype functions in high dimension are then given in Section 6. Some concluding remarks and open research directions are reported at the end of the manuscript.

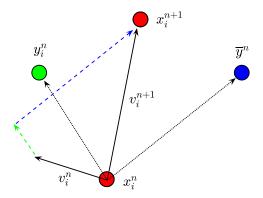


Fig. 1: Particle dynamics in the standard PSO model (2.2). Green and blue dashed arrows denote the influence of the local best and global best, respectively.

2. Second order stochastic models for particle swarm optimization

In the sequel we consider the following optimization problem

$$x^* \in \underset{x \in \mathbb{R}^d}{\operatorname{arg \, min}} \mathcal{F}(x) \,, \tag{2.1}$$

where $\mathcal{F}(x): \mathbb{R}^d \to \mathbb{R}$ is a given high dimensional objective function, which we wish to minimize. In machine learning the objective function allows the algorithm designer to encode the appropriate and expected behavior for the machine learning model, such as fitting well to the training data versus some loss function. Modern applications frequently require learning algorithms to operate in extremely high dimensional spaces [81,82]. In other applications, the natural objective of the learning task is a possibly nonsmooth and non-convex function [83]. Common examples include training deep neural networks and tensor decomposition problems. In contrast to gradient based optimizers and other metaheuristic solvers, PSO solve the minimization problem (2.1) by starting from a population of candidate solutions, represented by particles, and moving these particles in the search space according to simple mathematical relationships on particle position and speed. The movement of each particle is influenced by its best known local position, but it is also driven to the best collective position of the swarm in the search space, which is updated when the particles find better positions (see Figure 1).

2.1. The standard PSO algorithm

The method is based on introducing N particles with position $x_i \in \mathbb{R}^d$ and speed $v_i \in \mathbb{R}^d$, i = 1, ..., N. In the *standard PSO algorithm* the particle positions and velocities, starting with an initial x_i^0 and v_i^0 assigned, are updated according to the following rule [1]

$$x_i^{n+1} = x_i^n + v_i^{n+1},$$

$$v_i^{n+1} = v_i^n + c_1 R_1^n (y_i^n - x_i^n) + c_2 R_2^n (\overline{y}^n - x_i^n),$$
(2.2)

where the values $c_1, c_2 \in \mathbb{R}$ are the acceleration coefficients, y_i^n is the local best position found by the i particle up to that iteration, and \overline{y}^n is the global best position found among all the particles up to that iteration. The terms R_1^n and R_2^n denote two d-dimensional diagonal matrices with random numbers uniformly distributed in [0,1] on their diagonals. These numbers are generated at each iteration and for each particle. Typically, the values of x_i and v_i are restricted within a specific search domain $X = [X_{min}, X_{max}]^d$ and velocity range $V = [-V_{max}, V_{max}]^d$. Different boundary conditions are usually applied in the search space X.

The local best y_i^n and global best \overline{y}^n are defined by the following relationships

$$y_{i}^{0} = x_{i}^{0},$$

$$y_{i}^{n+1} = \begin{cases} y_{i}^{n} & \text{if } \mathcal{F}(x_{i}^{n+1}) \geq \mathcal{F}(y_{i}^{n}), \\ x_{i}^{n+1} & \text{if } \mathcal{F}(x_{i}^{n+1}) < \mathcal{F}(y_{i}^{n}), \end{cases}$$

$$\overline{y}^{0} = \operatorname{argmin}\{\mathcal{F}(x_{1}^{0}), \mathcal{F}(x_{2}^{0}), \dots, \mathcal{F}(x_{N}^{0})\},$$

$$\overline{y}^{n+1} = \operatorname{argmin}\{\mathcal{F}(y_{1}^{n+1}), \mathcal{F}(y_{2}^{n+1}), \dots, \mathcal{F}(y_{N}^{n+1})\}.$$
(2.3)

Another way to represent the local best, which will be useful in the sequel, is the following [6]

$$y_i^{n+1} = y_i^n + \frac{1}{2} \left(x_i^{n+1} - y_i^n \right) S(x_i^{n+1}, y_i^n), \tag{2.4}$$

where

$$S(x,y) = (1 + \operatorname{sgn}(\mathcal{F}(y) - \mathcal{F}(x))). \tag{2.5}$$

2.2. The stochastic differential PSO system

In order to derive a time continuous version of the PSO algorithm (2.2), we rewrite it in the form

$$x_{i}^{n+1} = x_{i}^{n} + v_{i}^{n+1},$$

$$v_{i}^{n+1} = v_{i}^{n} + \frac{c_{1}}{2} (y_{i}^{n} - x_{i}^{n}) + \frac{c_{2}}{2} (\overline{y}^{n} - x_{i}^{n}) + \frac{c_{1}}{2} \widetilde{R}_{1} (y_{i}^{n} - x_{i}^{n}) + \frac{c_{2}}{2} \widetilde{R}_{2} (\overline{y}^{n} - x_{i}^{n}),$$

$$(2.6)$$

where $\widetilde{R}_k = (2R_k - 1)$, k = 1, 2. We can interpret (2.6) as a semi-implicit time discretization method for SDEs with time stepping $\Delta t = 1$ where the implicit Euler scheme has been used for the first equation and the Euler-Maruyama method is used for the second one. Note that, the particular distribution of the random noise will not change the corresponding stochastic differential system provided the noise has the same mean value and variance. In the case of the PSO model (2.6), since the random terms are uniformly distributed in [-1,1], the mean value is 0 and the corresponding variance is 1/3.

We can then write the time continuous formulation as a second order system of SDEs in Itô form defining the *stochastic differential PSO system*

$$\begin{split} dX_{t}^{i} &= V_{t}^{i} dt, \\ dV_{t}^{i} &= \lambda_{1} \left(Y_{t}^{i} - X_{t}^{i} \right) dt + \lambda_{2} \left(\overline{Y}_{t} - X_{t}^{i} \right) dt \\ &+ \sigma_{1} D(Y_{t}^{i} - X_{t}^{i}) dB_{t}^{1,i} + \sigma_{2} D(\overline{Y}_{t} - X_{t}^{i}) dB_{t}^{2,i}, \end{split}$$
(2.7)

with

$$\lambda_k = \frac{c_k}{2}, \quad \sigma_k = \frac{c_k}{2\sqrt{3}}, \quad k = 1, 2$$
 (2.8)

the drift and diffusion coefficients and

$$D(X_t) = \operatorname{diag}\{(X_t)_1, (X_t)_2, \dots, (X_t)_d\},$$
(2.9)

a d-dimensional diagonal matrix.

In (2.7) the vectors $B_t^k = \left((B_t^k)_1, (B_t^k)_2, \dots, (B_t^k)_d\right)^T$, k = 1, 2 denote d independent 1-dimensional $Brownian\ motions$ and depend on the i-th particle. One critical aspect is the definition of the best positions Y_t^i and \overline{Y}_t which in the PSO method make use of the past history of the particles. Thanks to (2.4), for a positive constant ν , we can approximate y_i^{n+1} with

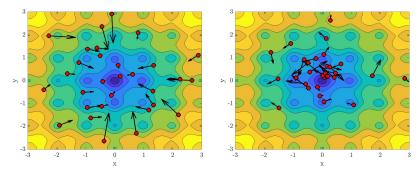


Fig. 2: Snapshots of the PSO minimization process (2.12) for the two-dimensional Ackley function (see Table 5) using N=30 particles, with m=0, $c_1=0.25$ and $c_2=2$.

the following differential system for the local best

$$dY_t^i = \nu \left(X_t^i - Y_t^i \right) S(X_t^i, Y_t^i) dt, \tag{2.10}$$

with $Y_0^i=X_0^i$ and consequently define

$$\overline{Y}_t = \operatorname{argmin} \left\{ \mathcal{F}(Y_t^1), \mathcal{F}(Y_t^2), \dots, \mathcal{F}(Y_t^N) \right\}. \tag{2.11}$$

Note that, equation (2.10) does not describe the evolution of the local best, but rather a time continuous approximation of its evolution.

2.3. Stochastic differential PSO model with inertia

To optimize the search algorithm, the value $c_k=2, k=1,2$ was adopted in early PSO research. This value, which corresponds to $\lambda_k=1$ and $\sigma_k=1/\sqrt{3}, k=1,2$ in the SDEs form, however, may lead to unstable dynamics with particle speed increase without control. The use of hard bounds on velocity in $[-V_{\rm max},V_{\rm max}]^d$ is one way to control the velocities. However, the value of $V_{\rm max}$ is problem-specific and difficult to determine. For this reason, the *PSO algorithm with inertia* has been considered [5]

$$x_i^{n+1} = x_i^n + v_i^{n+1}, v_i^{n+1} = mv_i^n + c_1 R_1^n (y_i^n - x_i^n) + c_2 R_2^n (\overline{y}^n - x_i^n),$$
(2.12)

where $m \in (0,1]$ is the *inertia weight*. The above system can be rewritten as

$$x_i^{n+1} = x_i^n + v_i^{n+1},$$

$$mv_i^{n+1} = mv_i^n - (1-m)v_i^{n+1} + c_1R_1^n (y_i^n - x_i^n) + c_2R_2^n (\overline{y}^n - x_i^n).$$
(2.13)

In this case, we can interpret the second equation as a semi-implicit Euler-Maruyama method, that is implicit in v_i and explicit in x_i , hence the corresponding stochastic differential PSO system with inertia reads

$$dX_t^i = V_t^i dt,$$

$$mdV_t^i = -\gamma V_t^i dt + \lambda_1 \left(Y_t^i - X_t^i \right) dt + \lambda_2 \left(\overline{Y}_t - X_t^i \right) dt + \sigma_1 D(Y_t^i - X_t^i) dB_t^{1,i} + \sigma_2 D(\overline{Y}_t - X_t^i) dB_t^{2,i},$$

$$(2.14)$$

where $\gamma=(1-m)\geq 0$ is the friction coefficient. Thus, the constant γ acts effectively as a friction coefficient, and can be related to the fluidity of the medium in which particles move. System (2.14) is reminescent of other second order stochastic particle system with inertia [73,74]. However, note that here, the inertia weight m and the friction coefficient γ are not independent.

In practice, in the PSO method (2.12) the parameter γ is often initially set to some low value, which corresponds to a system where particles move in a low viscosity medium and perform extensive exploration, and gradually increased to a higher value closer to one, where the system is more dissipative and would more easily concentrate into local minima. Most PSO approaches, nowadays, are based on (2.12) (or some variant) which is usually referred to as canonical PSO method to distinguish it from the original PSO method (2.2) (see [4]). Similarly we will refer to (2.7)-(2.10) as the original stochastic differential PSO (SD-PSO) system and to (2.14)-(2.10) as the canonical SD-PSO system.

Remark 2.1: We underline that the PSO stochastic systems (2.14)-(2.10) if discretized properly yields the PSO algorithm with inertia (2.12). This is achieved discretizing (2.14) implicitly in V_t^i and explicitly in X_t^i , and (2.10) implicitly in X_t^i and explicitly in Y_t^i . Taking $\Delta t = 1$, $\nu = 1/2$, the drift and diffusion terms satisfying (2.8), and a uniform noise permits to recover exactly (2.12). We refer to the last part of the manuscript containing the numerical examples for further details.

3. Mean-field particle swarm optimization

In this section we introduce a modified version of the canonical stochastic differential PSO system for which we can formally compute its mean-field limit. We first consider the case in absence of memory effects and then we extend the results to the general case. Throughout this note, our theoretical analysis assumes the cost function \mathcal{F} satisfies the following

Assumption 1: For the given cost function $\mathcal{F}: \mathbb{R}^d \to \mathbb{R}$, it holds that:

- (1) There exists some constant L > 0 such $|\mathcal{F}(x) \mathcal{F}(y)| \le L(|x| + |y|)|x y|$ for all $x, y \in \mathbb{R}^d$;
- (2) \mathcal{F} is bounded from below with $-\infty < \underline{\mathcal{F}} := \inf \mathcal{F}$ and there exists some constant $C_u > 0$ such that

$$\mathcal{F}(x) - \underline{\mathcal{F}} \leq C_u(1+|x|^2)$$
 for all $x \in \mathbb{R}^d$;

(3) \mathcal{F} has quadratic growth at infinity. Namely, there exist constants C_l , M > 0 such that

$$\mathcal{F}(x) - \mathcal{F} \ge C_l |x|^2$$
 for all $|x| \ge M$.

3.1. Regularized PSO dynamics without memory effects

To simplify the mathematical description, let us consider a PSO approach where the dynamic is instantaneous without memory of the local best positions and the global best has been regularized as in [50]. The corresponding second order system of SDEs describing the *regularized SD-PSO dynamics* takes the form^a

$$dX_t^{i,N} = V_t^{i,N} dt,$$

$$mdV_t^{i,N} = -\gamma V_t^{i,N} dt + \lambda (X^{\alpha}(\rho_t^N) - X_t^{i,N}) dt$$

$$+ \sigma D(X^{\alpha}(\rho_t^N) - X_t^{i,N}) dB_t^i,$$
(3.1)

where the \mathbb{R}^d -valued functions $X^{i,N}_t$ and $V^{i,N}_t$ denote the position and velocity of the i-th particle at time t, and $\{(B^i_t)_{t\geq 0}\}_{i=1}^N$ are N independent d-dimensional Brownian motions. Here the weighted average regularization of the global best is given by

$$X^{\alpha}(\rho_t^N) := \frac{\int_{\mathbb{R}^d} x \omega_{\alpha}^{\mathcal{F}}(x) \rho_t^N(dx)}{\int_{\mathbb{R}^d} \omega_{\alpha}^{\mathcal{F}}(x) \rho_t^N(dx)},$$
(3.2)

with the empirical measure $\rho^N := \frac{1}{N} \sum_{i=1}^N \delta_{X^{i,N}}$, which is the spacial marginal of $f^N := \frac{1}{N} \sum_{i=1}^N \delta_{(X^{i,N},V^{i,N})}$. The choice of the weight function $\omega_{\alpha}^{\mathcal{F}}(x) := e^{-\alpha \mathcal{F}(x)}$ in (3.2) comes from the well-known *Laplace principle*, a

a The superscript N is used to emphasize the dependence on the number of particles in the system.

classical result in large deviation theory, which states that for any probability measure $\rho \in \mathcal{P}(\mathbb{R}^d)$ compactly supported, it holds

$$\lim_{\alpha \to \infty} \left(-\frac{1}{\alpha} \log \left(\int_{\mathbb{R}^d} e^{-\alpha \mathcal{F}(x)} \rho(dx) \right) \right) = \inf_{x \in \text{supp}(\rho)} \mathcal{F}(x). \tag{3.3}$$

Therefore, for large values of $\alpha \gg 1$ the regularized global best $X^{\alpha}(\rho_t^N) \approx X_t^*$, where

$$X_t^* = \operatorname{argmin} \left\{ \mathcal{F}(X_t^{1,N}), \mathcal{F}(X_t^{2,N}), \dots, \mathcal{F}(X_t^{N,N}) \right\}.$$

We emphasize that the stochastic particle system (3.1) has locally Lipschitz coefficients, thus it admits strong solutions and pathwise uniqueness holds up to any finite time T > 0, see [18,84]. The above system of SDEs in the sequel is considered in a general setting, without necessarily satisfying the PSO constraint (2.8).

As the particle number $N \to \infty$, one expects to derive the mean-filed PSO description without local best characterized by the following nonlinear Vlasov-Fokker-Planck equation

$$\partial_t f + v \cdot \nabla_x f = \nabla_v \cdot \left(\frac{\gamma}{m} v f + \frac{\lambda}{m} (x - X^{\alpha}(\rho)) f + \frac{\sigma^2}{2m^2} D(x - X^{\alpha}(\rho))^2 \nabla_v f \right)$$
(3.4)

where we have used the identity

$$\sum_{j=1}^{d} \frac{\partial^2}{\partial v_j^2} \left((x - X^{\alpha}(\rho))_j^2 f \right) = \nabla_v \cdot \left(D(x - X^{\alpha}(\rho))^2 \nabla_v f \right)$$
 (3.5)

with $D(x - X^{\alpha}(\rho))^2$ the diagonal matrix given by the square of $D(x - X^{\alpha}(\rho))$. Equation (3.4) represents the mean-field PSO (MF-PSO) model without local best and should be accompanied by initial (and boundary) data, and normalization

$$\iint_{\mathbb{R}^{2d}} f(t, dx, dv) = 1.$$

We refer to [13,43–45,49] and the references therein, for more details and rigorous results about mean-field models of Vlasov-Fokker-Planck type. Note, however, that the presence of $X^{\alpha}(\rho)$ makes the Vlasov-Fokker-Planck equation nonlinear and nonlocal. This is nonstandard in the literature and raises several analytical and numerical questions (see [51,61]).

3.1.1. Mean-field limit

In this section, following [70] we provide a rigorous justification of the mean-field limit of PSO model (3.1) towards its mean-field PDE (3.4) through a compactness argument. More precisely, we first prove that the sequence of empirical measures $\{f^N\}_{N\geq 2}$ ($f^N=\frac{1}{N}\sum_{i=1}^N \delta_{(X^{i,N},V^{i,N})}$ are $\mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d)\times\mathcal{C}([0,T];\mathbb{R}^d))$ -valued random variables) is tight. Prokhorov's theorem indicates that there exists a subsequence of $\{f^N\}_{N\geq 2}$ converging in law to a random measure f. Then, to identify the limit, we verify that the limit measure f is a weak solution to the mean-field PSO equation (3.4) almost surely, while the uniqueness of the weak solution to PDE (3.4) yields that f is actually deterministic. Our main result can be described in the following way:

Theorem 1: Let \mathcal{F} satisfy Assumption 1 and $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$. For any $N \geq 2$, we assume that $\{(X_t^{i,N}, V_t^{i,N})_{t \in [0,T]}\}_{i=1}^N$ is the unique solution to the SD-PSO system (3.1) with $f_0^{\otimes N}$ -distributed initial data $\{X_0^{i,N}, V_0^{i,N}\}_{i=1}^N$. Then the limit (denoted by f) of the sequence of the empirical measure $f^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X_i^{i,N}, V_i^{i,N})}$ exists. Moreover, f is the unique weak solution to the MF-PSO equation (3.4).

To obtain the above theorem, let us first prove the following lemma on a uniform moment estimate for the particle system (3.1).

Lemma 3.1: Let \mathcal{F} satisfy Assumption 1 and $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$. For any $N \geq 2$, assume that $\{(X_t^{i,N}, V_t^{i,N})_{t \in [0,T]}\}_{i=1}^N$ is the unique solution to the SD-PSO system (3.1) with $f_0^{\otimes N}$ -distributed initial data $\{(X_0^{i,N}, V_0^{i,N})\}_{i=1}^N$. Then there exists a constant K > 0 independent of N such that

$$\sup_{i=1,\dots,N} \left\{ \sup_{t \in [0,T]} \mathbb{E} \left[|X_t^{i,N}|^2 + |X_t^{i,N}|^4 + |V_t^{i,N}|^2 + |V_t^{i,N}|^4 \right] \right\} \\
+ \sup_{t \in [0,T]} \mathbb{E} \left[|X^{\alpha}(\rho_t^N)|^2 + |X^{\alpha}(\rho_t^N)|^4 \right] \le K.$$
(3.6)

The proof follows similar arguments as in [51, Lemma 3.4].

We treat $(X^{i,N}, V^{i,N}): \Omega \mapsto \mathcal{C}([0,T]; \mathbb{R}^d) \times \mathcal{C}([0,T]; \mathbb{R}^d)$. Then $f^N = \sum_{i=1}^N \delta_{(X^{i,N},V^{i,N})}: \Omega \mapsto \mathcal{P}(\mathcal{C}([0,T]; \mathbb{R}^d) \times \mathcal{C}([0,T]; \mathbb{R}^d))$ is a random measure. Let us denote $\mathcal{L}(f^N) := \operatorname{Law}(f^N) \in \mathcal{P}(\mathcal{P}(\mathcal{C}([0,T]; \mathbb{R}^d) \times \mathcal{C}([0,T]; \mathbb{R}^d)))$. We can prove that $\{\mathcal{L}(f^N)\}_{N\geq 2}$ is tight, or we say $\{f^N\}_{N\geq 2}$ is tight, which can be done by verifying the Aldous criteria [85] as presented below:

Lemma 3.2: Let $\{X^n\}_{n\in\mathbb{N}}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and valued in $\mathcal{C}([0,T];\mathbb{R}^d)$. The sequence of probability distributions $\{\mu_{X^n}\}_{n\in\mathbb{N}}$ of $\{X^n\}_{n\in\mathbb{N}}$ is tight on $\mathcal{C}([0,T];\mathbb{R}^d)$ if the following two conditions hold.

(Con1) For all $t \in [0,T]$, the set of distributions of X_t^n , denoted by $\{\mu_{X_t^n}\}_{n \in \mathbb{N}}$, is tight as a sequence of probability measures on \mathbb{R}^d .

(Con2) For all $\varepsilon > 0$, $\eta > 0$, there exists $\delta_0 > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ and for all discrete-valued $\sigma(X_s^n; s \in [0, T])$ -stopping times β with $0 \leq \beta + \delta_0 \leq T$, it holds that

$$\sup_{\delta \in [0,\delta_0]} \mathbb{P}\left(|X_{\beta+\delta}^n - X_{\beta}^n| \ge \eta \right) \le \varepsilon. \tag{3.7}$$

We can then prove:

Theorem 3.3: Let \mathcal{F} satisfy Assumption 1 and $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$. For any $N \geq 2$, we assume that $\{(X_t^{i,N}, V_t^{i,N})_{t \in [0,T]}\}_{i=1}^N$ is the unique solution to the SD-PSO system (3.1) with $f_0^{\otimes N}$ -distributed initial data $\{X_0^{i,N}, V_0^{i,N}\}_{i=1}^N$. Then the sequence $\{\mathcal{L}(f^N)\}_{N\geq 2}$ is tight in $\mathcal{P}(\mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d) \times \mathcal{C}([0,T];\mathbb{R}^d)))$.

Proof: According to Proposition 2.2 (ii) in [43, Proposition 2.2 (ii)], we only need to prove that $\{\mathcal{L}((X^{1,N},V^{1,N}))\}_{N\geq 2}$ is tight in $\mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d)\times\mathcal{C}([0,T];\mathbb{R}^d))$ because of the exchangeability of the particle system. It is sufficient to justify conditions (Con1) and (Con2) in Lemma 3.2.

• Step 1: Checking (Con1). For any $\varepsilon > 0$, there exists a compact subset $U_{\varepsilon} := \{(x,v): \ |x|^2 + |v|^2 \leq \frac{K}{\varepsilon} \}$ such that by Markov's inequality

$$\begin{split} \mathcal{L}((\boldsymbol{X}_t^{1,N}, \boldsymbol{V}_t^{1,N})) \ \left((\boldsymbol{U}_{\varepsilon})^c \right) &= \mathbb{P}\left(|\boldsymbol{X}_t^{1,N}|^2 + |\boldsymbol{V}_t^{1,N}|^2 > \frac{K}{\varepsilon} \right) \\ &\leq \frac{\varepsilon \mathbb{E}[|\boldsymbol{X}_t^{1,N}|^2 + |\boldsymbol{V}_t^{1,N}|^2]}{K} \leq \varepsilon, \quad \forall \ N \geq 2 \,, \end{split}$$

where we have used Lemma 3.1 in the last inequality. This means that for each $t \in [0,T]$, the sequence $\{\mathcal{L}((X_t^{1,N},V_t^{1,N}))\}_{N\geq 2}$ is tight, which verifies condition (Con1) in Lemma 3.2.

• Step 2: Checking (Con2). Let β be a $\sigma((X_s^{1,N}, V_s^{1,N}); s \in [0,T])$ stopping time with discrete values such that $\beta + \delta_0 \leq T$. It is easy to see that

$$\mathbb{E}[|X_{\beta+\delta}^{1,N} - X_{\beta}^{1,N}|^2] \le \delta \int_0^T \mathbb{E}[|V_s^{1,N}|^2] ds \le C\delta, \tag{3.8}$$

where C > 0 is independent of N by (3.6). Furthermore,

$$\begin{split} V_{\beta+\delta}^{1,N} - V_{\beta}^{1,N} &= -\frac{\gamma}{m} \int_{\beta}^{\beta+\delta} V_s^{1,N} ds + \frac{\lambda}{m} \int_{\beta}^{\beta+\delta} (X^{\alpha}(\rho_s^N) - X_s^{1,N}) ds \\ &+ \frac{\sigma}{m} \int_{\beta}^{\beta+\delta} D(X^{\alpha}(\rho_s^N) - X_s^{1,N}) dB_s^1 \,. \end{split}$$

Notice that

$$\mathbb{E}\left[\left|\int_{\beta}^{\beta+\delta} (X^{\alpha}(\rho_s^N) - X_s^{1,N}) ds\right|^2\right] \le \delta \int_0^T \mathbb{E}\left[|X^{\alpha}(\rho_s^N) - X_s^{1,N}|^2\right] ds$$

$$\le 2\delta T \left(\sup_{t \in [0,T]} \mathbb{E}\left[|X_t^{1,N}|^2\right] + \sup_{t \in [0,T]} \mathbb{E}\left[|X^{\alpha}(\rho_t^N)|^2\right]\right) \le 2TK\delta,$$
(3.9)

where we have used Lemma 3.1 in the last inequality. Similarly we have

$$\mathbb{E}\left[\left|\int_{\beta}^{\beta+\delta} V_s^{1,N} ds\right|^2\right] \le TK\delta. \tag{3.10}$$

Further we apply Itô's isometry

$$\mathbb{E}\left[\left|\int_{\beta}^{\beta+\delta} D(X^{\alpha}(\rho_{s}^{N}) - X_{s}^{1,N})dB_{s}^{1}\right|^{2}\right] = \mathbb{E}\left[\int_{\beta}^{\beta+\delta} |X^{\alpha}(\rho_{s}^{N}) - X_{s}^{1,N}|^{2}ds\right] \\
\leq \delta^{\frac{1}{2}} \mathbb{E}\left[\left(\int_{0}^{T} |X^{\alpha}(\rho_{s}^{N}) - X_{s}^{1,N}|^{4}ds\right)^{\frac{1}{2}}\right] \\
\leq \delta^{\frac{1}{2}}\left(\int_{0}^{T} \mathbb{E}[|X^{\alpha}(\rho_{s}^{N}) - X_{s}^{1,N}|^{4}]ds\right)^{\frac{1}{2}} \leq \delta^{\frac{1}{2}}T^{\frac{1}{2}}(8K)^{\frac{1}{2}}. \tag{3.11}$$

Combining estimates (3.9)–(3.11) one has

$$\mathbb{E}[|V_{\beta+\delta}^{1,N} - V_{\beta}^{1,N}|^2] \le C(\gamma.\lambda, m, \sigma, T, K) \left(\delta^{\frac{1}{2}} + \delta\right). \tag{3.12}$$

Hence, for any $\varepsilon > 0$, $\eta > 0$, there exists some $\delta_0 > 0$ such that for all $N \geq 2$ it holds that

$$\sup_{\delta \in [0,\delta_{0}]} \mathbb{P}\left(|X_{\beta+\delta}^{1,N} - X_{\beta}^{1,N}|^{2} + |V_{\beta+\delta}^{1,N} - V_{\beta}^{1,N}|^{2} \ge \eta \right) \\
\leq \sup_{\delta \in [0,\delta_{0}]} \frac{\mathbb{E}\left[|X_{\beta+\delta}^{1,N} - X_{\beta}^{1,N}|^{2} + |V_{\beta+\delta}^{1,N} - V_{\beta}^{1,N}|^{2} \right]}{\eta} \le \varepsilon.$$
(3.13)

Hence (Con2) is verified.

For any $\varphi \in \mathcal{C}^2_c(\mathbb{R}^d \times \mathbb{R}^d)$, define a functional on $\mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d) \times \mathcal{C}([0,T];\mathbb{R}^d))$ as follows

$$F_{\varphi}(f) := \langle \varphi(\mathbf{x}_{t}, \mathbf{v}_{t}), f(d\mathbf{x}, d\mathbf{v}) \rangle - \langle \varphi(\mathbf{x}_{0}, \mathbf{v}_{0}), f(d\mathbf{x}, d\mathbf{v}) \rangle + \int_{0}^{t} \langle \mathbf{v}_{s} \cdot \nabla_{x} \varphi, f(d\mathbf{x}, d\mathbf{v}) \rangle ds$$

$$- \frac{\gamma}{m} \int_{0}^{t} \langle \mathbf{v}_{s} \cdot \nabla_{v} \varphi, f(d\mathbf{x}, d\mathbf{v}) \rangle ds + \frac{\lambda}{m} \int_{0}^{t} \langle (\mathbf{x}_{s} - X^{\alpha}(\rho_{s})) \cdot \nabla_{v} \varphi, f(d\mathbf{x}, d\mathbf{v}) \rangle ds$$

$$- \frac{\sigma^{2}}{2m^{2}} \int_{0}^{t} \sum_{k=1}^{d} \langle (\mathbf{x}_{s} - X^{\alpha}(\rho_{s}))_{k}^{2} \frac{\partial^{2} \varphi}{\partial v_{k}^{2}}, f(d\mathbf{x}, d\mathbf{v}) \rangle ds$$

$$= \langle \varphi(x, v), f_{t}(dx, dv) \rangle - \langle \varphi(x, v), f_{0}(dx, dv) \rangle + \int_{0}^{t} \langle v \cdot \nabla_{x} \varphi, f_{s}(dx, dv) \rangle ds$$

$$- \frac{\gamma}{m} \int_{0}^{t} \langle v \cdot \nabla_{v} \varphi, f_{s}(dx, dv) \rangle ds + \frac{\lambda}{m} \int_{0}^{t} \langle (x - X^{\alpha}(\rho_{s})) \cdot \nabla_{v} \varphi, f_{s}(dx, dv) \rangle ds$$

$$- \frac{\sigma^{2}}{2m^{2}} \int_{0}^{t} \sum_{k=1}^{d} \langle (x - X^{\alpha}(\rho_{s}))_{k}^{2} \frac{\partial^{2} \varphi}{\partial v_{k}^{2}}, f_{s}(dx, dv) \rangle ds,$$

for all $f \in \mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d) \times \mathcal{C}([0,T];\mathbb{R}^d))$ and $\mathbf{x}, \mathbf{v} \in \mathcal{C}([0,T];\mathbb{R}^d)$, where $\rho_t(x) = \int_{\mathbb{R}^d} f_t(x, dv)$.

Then we have the following estimate by the reasoning in [70, Proposition 3.2].

Lemma 3.4: Let \mathcal{F} satisfy Assumption 1 and $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$. For any $N \geq 2$, assume that $\{(X_t^{i,N}, V_t^{i,N})_{t \in [0,T]}\}_{i=1}^N$ is the unique solution to the SD-PSO system (3.1) with $f_0^{\otimes N}$ -distributed initial data $\{(X_0^{i,N}, V_0^{i,N})\}_{i=1}^N$. There exists a constant C > 0 depending only on $\sigma, \gamma, \lambda, m, K, T$, and $\|\nabla \varphi\|_{\infty}$ such that

$$\mathbb{E}[|F_{\varphi}(f^N)|^2] \le \frac{C}{N},\tag{3.14}$$

where $f^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X^{i,N},V^{i,N})}$ is the empirical measure.

By Skorokhod's lemma (see [85, Theorem 6.7 on page 70]), using Theorem 3.3 we may find a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the processes $\{f^N\}_{N\in\mathbb{N}}$ converge to some process f as a random variable valued in $\mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d)) \times \mathcal{C}([0,T];\mathbb{R}^d))$ almost surely. In particular, we have that for all $t \in [0,T]$ and $\phi \in C_b(\mathbb{R}^d \times \mathbb{R}^d)$,

$$\lim_{N \to \infty} |\langle \phi, f_t^N - f_t \rangle| + |X^{\alpha}(\rho_t^N) - X^{\alpha}(\rho_t)| = 0, \quad \text{a.s.}$$
 (3.15)

Indeed, according to Assumption 1, one has $xe^{-\alpha \mathcal{F}(x)}, e^{-\alpha \mathcal{F}(x)} \in \mathcal{C}_b(\mathbb{R}^d)$, which gives

$$\lim_{N \to \infty} X^{\alpha}(\rho_t^N) = \lim_{N \to \infty} \frac{\langle x e^{-\alpha \mathcal{F}(x)}, \rho_t^N(dx) \rangle}{\langle e^{-\alpha \mathcal{F}(x)}, \rho_t^N(dx) \rangle} = \frac{\langle x e^{-\alpha \mathcal{F}(x)}, \rho_t(dx) \rangle}{\langle e^{-\alpha \mathcal{F}(x)}, \rho_t(dx) \rangle} = X^{\alpha}(\rho_t) \quad \text{a.s.}$$

Lemma 3.5: [51, Lemma 3.3] Let \mathcal{F} satisfy Assumption 1 and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Then it holds that

$$|X^{\alpha}(\mu)|^2 \le b_1 + b_2 \int_{\mathbb{R}^d} |x|^2 \mu(dx),$$
 (3.16)

where b_1 and b_2 depends only on M, C_u , and C_l .

For each A > 0, it follows from (3.15) that

$$\mathbb{E}\left[\iint_{\mathbb{R}^{2d}} ((|x|^4 + |v|^4) \wedge A) f_t(dx, dv)\right]$$

$$= \mathbb{E}\left[\lim_{N \to \infty} \iint_{\mathbb{R}^{2d}} ((|x|^4 + |v|^4) \wedge A) f_t^N(dx, dv)\right]$$

$$\leq \lim_{N \to \infty} \frac{\sum_{i=1}^N \mathbb{E}[|X_t^{i,N}|^4 + |V_t^{i,N}|^4]}{N} \leq K,$$

where we have used Lemma 3.1. Letting $A \to \infty$, we have

$$\sup_{t \in [0,T]} \mathbb{E} \left[\iint_{\mathbb{R}^{2d}} (|x|^4 + |v|^4) f_t(dx, dv) \right] \le K.$$
 (3.17)

Then Lemma 3.5 implies that

$$\mathbb{E}[|X^{\alpha}(\rho_t)|^4] < \infty, \tag{3.18}$$

for all $t \in [0, T]$. Furthermore, it holds that

$$\lim_{N \to \infty} \mathbb{E}\left[\left|\langle \phi, f_t^N - f_t \rangle\right|^2 + \left|X^{\alpha}(\rho_t^N) - X^{\alpha}(\rho_t)\right|^2\right] = 0, \tag{3.19}$$

which follows directly from the pointwise convergences of $\langle \phi, f_t^N - f_t \rangle$ and $X^{\alpha}(\rho_t^N) - X^{\alpha}(\rho_t)$, and the uniform estimate (3.6) in Lemma 3.1 and (3.18). We can now prove the main result in Theorem 1:

Proof: (Theorem 1) Suppose the $\mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d) \times \mathcal{C}([0,T];\mathbb{R}^d))$ -valued random variable f is the limit of a subsequence of the empirical measure $f^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X^{i,N},V^{i,N})}$. W.l.o.g., Denote the subsequence by itself. We may continue to work on the above common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ by Skorokhod's lemma where the convergence is holding almost surely (see

(3.15) for instance). We may first check that f_t is a.s. continuous in time. Indeed for any $\phi \in \mathcal{C}_b(\mathbb{R}^{2d})$ and $t_n \to t$ we may apply dominated convergence theorem

$$\iint_{\mathcal{C}([0,T];\mathbb{R}^d)\times\mathcal{C}([0,T];\mathbb{R}^d)} \phi(\mathbf{x}_{t_n}, \mathbf{v}_{t_n}) f(d\mathbf{x}, d\mathbf{v})$$

$$\to \iint_{\mathcal{C}([0,T];\mathbb{R}^d)\times\mathcal{C}([0,T];\mathbb{R}^d)} \phi(\mathbf{x}_t, \mathbf{v}_t) f(d\mathbf{x}, d\mathbf{v}) \quad \text{a.s.},$$

which gives

$$\iint_{\mathbb{R}^{2d}} \phi(x, v) f_{t_n}(dx, dv) \to \iint_{\mathbb{R}^{2d}} \phi(x, v) f_t(dx, dv) \quad \text{a.s.}$$

For $\varphi \in \mathcal{C}^2_c(\mathbb{R}^{2d})$, using the convergence result in (3.19) one has

$$\lim_{N \to \infty} \mathbb{E}\left[\left| \left(\langle \varphi, f_t^N \rangle - \langle \varphi, f_0^N \rangle \right) - \left(\langle \varphi, f_t \rangle - \langle \varphi, f_0 \rangle \right) \right| \right] = 0. \tag{3.20}$$

Further we notice that

$$\begin{split} &\left| \int_0^t \langle (x - X^{\alpha}(\rho_s^N)) \cdot \nabla_v \varphi, f_s^N \rangle ds - \int_0^t \langle (x - X^{\alpha}(\rho_s)) \cdot \nabla_v \varphi, f_s \rangle ds \right| \\ & \leq \int_0^t \left| \langle (x - X^{\alpha}(\rho_s^N)) \cdot \nabla_v \varphi, f_s^N - f_s \rangle \right| ds \\ & + \int_0^t \left| \langle (X^{\alpha}(\rho_s) - X^{\alpha}(\rho_s^N)) \cdot \nabla_v \varphi, f_s \rangle \right| ds \\ & =: \int_0^t \left| I_1^N(s) | ds + \int_0^t \left| I_2^N(s) | ds \right|. \end{split}$$

One computes

$$\begin{split} & \mathbb{E}[|I_1^N(s)|] \\ & \leq \mathbb{E}[|\langle x \cdot \nabla_v \varphi, f_s^N - f_s \rangle|] + \mathbb{E}[|X^{\alpha}(\rho_s^N) \cdot \langle \nabla_v \varphi, f_s^N - f_s \rangle|] \\ & \leq \mathbb{E}[|\langle x \cdot \nabla_v \varphi, f_s^N - f_s \rangle|] + K^{\frac{1}{2}} (\mathbb{E}[|\langle \nabla_v \varphi, f_s^N - f_s \rangle|^2])^{\frac{1}{2}} \,, \end{split}$$

where we have used Lemma 3.1 in the second inequality. Since φ has a compact support, applying (3.19) leads to

$$\lim_{N \to \infty} \mathbb{E}\left[|I_1^N(s)| \right] = 0. \tag{3.21}$$

Moreover, the uniform boundedness of $\mathbb{E}[|I_1^N(s)|]$ follows directly from (3.17), (3.18), and the estimates in Lemma 3.1, which by the dominated convergence theorem implies

$$\lim_{N \to \infty} \int_0^t \mathbb{E}[|I_1^N(s)|] ds = 0.$$
 (3.22)

As for I_2^N , we know that

$$\left| \langle (X^{\alpha}(\rho_s) - X^{\alpha}(\rho_s^N)) \cdot \nabla_v \varphi, f_s \rangle \right| \le \|\nabla_v \varphi\|_{\infty} |X^{\alpha}(\rho_s) - X^{\alpha}(\rho_s^N)|. \quad (3.23)$$

Hence by (3.19) it yields that

$$\lim_{N \to \infty} \mathbb{E}[|I_2^N(s)|] = 0. \tag{3.24}$$

Again by the dominated convergence theorem, we have

$$\lim_{N \to \infty} \int_0^t \mathbb{E}[|I_2^N(s)|] ds = 0.$$
 (3.25)

This combined with (3.22) leads to

$$\lim_{N \to \infty} \mathbb{E}\left[\left| \int_0^t \langle (x - X^{\alpha}(\rho_s^N)) \cdot \nabla_v \varphi, f_s^N \rangle ds - \int_0^t \langle (x - X^{\alpha}(\rho_s)) \cdot \nabla_v \varphi, f_s \rangle ds \right|\right] = 0.$$
(3.26)

Similarly we split the error

$$\begin{split} &\left| \int_0^t \langle (x - X^\alpha(\rho_s^N))_k^2 \frac{\partial^2}{\partial v_k^2} \varphi, f_s^N \rangle ds - \int_0^t \langle (x - X^\alpha(\rho_s))_k^2 \frac{\partial^2}{\partial v_k^2} \varphi, f_s \rangle ds \right| \\ & \leq \left| \int_0^t \langle (x - X^\alpha(\rho_s^N))_k^2 \frac{\partial^2}{\partial v_k^2} \varphi, f_s^N - f_s \rangle ds \right| \\ & + \left| \int_0^t \langle ((x - X^\alpha(\rho_s^N))_k^2 - (x - X^\alpha(\rho_s))_k^2) \frac{\partial^2}{\partial v_k^2} \varphi, f_s \rangle ds \right| \\ & =: \int_0^t |I_3^N(s)| ds + \int_0^t |I_4^N(s)| ds \,. \end{split}$$

Following the same argument as for I_1^N and I_2^N , one has

$$\lim_{N \to \infty} \int_0^t \mathbb{E}[|I_3^N(s)|] ds = 0 \text{ and } \lim_{N \to \infty} \int_0^t \mathbb{E}[|I_4^N(s)|] ds = 0.$$
 (3.27)

This implies that

$$\lim_{N \to \infty} \mathbb{E} \left[\left| \int_0^t \sum_{k=1}^d \langle (x - X^{\alpha}(\rho_s^N))_k^2 \frac{\partial^2}{\partial v_k^2} \varphi(x), f_s^N \rangle ds \right. \\ \left. - \int_0^t \sum_{k=1}^d \langle (x - X^{\alpha}(\rho_s))_k^2 \frac{\partial^2}{\partial v_k^2} \varphi(x), f_s \rangle ds \right| \right] = 0.$$
 (3.28)

Moreover it is easy to get

$$\lim_{N \to \infty} \mathbb{E}\left[\left| \int_0^t \langle v \cdot \nabla_x \varphi, f_s^N \rangle ds - \int_0^t \langle v \cdot \nabla_x \varphi, f_s \rangle ds \right| \right] = 0 \tag{3.29}$$

and

$$\lim_{N \to \infty} \mathbb{E}\left[\left| \int_0^t \langle v \cdot \nabla_v \varphi, f_s^N \rangle ds - \int_0^t \langle v \cdot \nabla_v \varphi, f_s \rangle ds \right| \right] = 0.$$
 (3.30)

Collecting estimates (3.20), (3.26), (3.28), (3.29) and (3.30) we have

$$\lim_{N \to \infty} \mathbb{E}[|F_{\varphi}(f^N) - F_{\varphi}(f)|] = 0. \tag{3.31}$$

Then we have

$$\begin{split} \mathbb{E}[|F_{\varphi}(f)|] &\leq \mathbb{E}[|F_{\varphi}(f^N) - F_{\varphi}(f)|] + \mathbb{E}[|F_{\varphi}(f^N)|] \\ &\leq \mathbb{E}[|F_{\varphi}(f^N) - F_{\varphi}(f)|] + \frac{C}{\sqrt{N}} \to 0 \quad \text{as } N \to \infty \,, \end{split}$$

where we have used Lemma 3.4 in the last inequality. This implies that

$$F_{\omega}(f) = 0$$
 a.s. (3.32)

In other words, it holds that

$$\langle \varphi(x,v), f_t(dx,dv) \rangle - \langle \varphi(x,v), f_0(dx,dv) \rangle + \int_0^t \langle v \cdot \nabla_x \varphi, f_s(dx,dv) \rangle ds$$
$$- \frac{\gamma}{m} \int_0^t \langle v \cdot \nabla_v \varphi, f_s(dx,dv) \rangle ds + \frac{\lambda}{m} \int_0^t \langle (x - X^{\alpha}(\rho_s)) \cdot \nabla_v \varphi, f_s(dx,dv) \rangle ds$$
$$- \frac{\sigma^2}{2m^2} \int_0^t \sum_{k=1}^d \langle (x - X^{\alpha}(\rho_s))_k^2 \frac{\partial^2 \varphi}{\partial v_k^2}, f_s(dx,dv) \rangle ds = 0,$$

for any $\varphi \in \mathcal{C}^2_c(\mathbb{R}^{2d})$.

Until now we have proved that f a.s. is a weak solution to PDE (3.4). Finally combining the uniqueness of weak solution to (3.4) (see for example in [49]) and the arbitrariness of the subsequence of $\{f^N\}_{N\geq 2}$, the (deterministic) weak solution f to PDE (3.4) must be the limit of the whole sequence $\{f^N\}_{N\geq 2}$. We completed the proof.

3.2. Regularized PSO dynamic with memory and local best

Next, we consider the second order system of SDEs corresponding to the $regularized\ SD\text{-}PSO\ method\ with\ local\ best$

$$dX_{t}^{i,N} = V_{t}^{i,N} dt,$$

$$dY_{t}^{i,N} = \nu \left(X_{t}^{i,N} - Y_{t}^{i,N} \right) S^{\beta}(X_{t}^{i,N}, Y_{t}^{i,N}) dt,$$

$$mdV_{t}^{i,N} = -\gamma V_{t}^{i,N} dt + \lambda_{1} \left(Y_{t}^{i,N} - X_{t}^{i,N} \right) dt$$

$$+ \lambda_{2} \left(Y^{\alpha}(\overline{\rho}_{t}^{N}) - X_{t}^{i,N} \right) dt$$

$$+ \sigma_{1} D(Y_{t}^{i,N} - X_{t}^{i,N}) dB_{t}^{1,i}$$

$$+ \sigma_{2} D(Y^{\alpha}(\overline{\rho}_{t}^{N}) - X_{t}^{i,N}) dB_{t}^{2,i},$$
(3.33)

where, similarly to the previous case, we introduced the following regularized $global\ best$

$$Y^{\alpha}(\overline{\rho}_{t}^{N}) := \frac{\int_{\mathbb{R}^{d}} y \omega_{\alpha}^{\mathcal{E}}(y) \overline{\rho}_{t}^{N}(dy)}{\int_{\mathbb{R}^{d}} \omega_{\alpha}^{\mathcal{E}}(y) \rho_{t}^{N}(dy)},$$
(3.34)

with the empirical measure $\overline{\rho}^N := \frac{1}{N} \sum_{i=1}^N \delta_{Y^{i,N}}$, which is the Y-marginal of $f^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X^{i,N},Y^{i,N},V^{i,N})}$.

Furthermore, in the right hand side of (3.33) we have replaced the $\operatorname{sgn}(x)$ function with a $\operatorname{sigmoid}$, for example the hyperbolic tangent $\tanh(\beta x)$ for $\beta \gg 1$, and consider

$$S^{\beta}(x,y) = 1 + \tanh\left(\beta(\mathcal{F}(y) - \mathcal{F}(x))\right). \tag{3.35}$$

Thanks to these regularizations, also the stochastic particle system (3.33) has locally Lipschitz coefficients and therefore it admits strong solutions and pathwise uniqueness holds for any finite time T>0. Even in this case, the system of SDEs (3.33) is generalized without restricting the search parameters to the PSO constraint (2.8).

In order to derive a mean-field description of system (3.33), we can follow the same arguments as in Section 3.1.1. The only difference is that we have an additional variable Y, which can be treated easily because of the regularity of the function S^{β} . Namely we can prove the tightness of the empirical measures $\{f^N\}_{N\geq 2}$ by verifying the Aldous criteria (Lemma 3.2). Then there exists a subsequence of $\{f^N\}_{N\geq 2}$ converging in law to a deterministic measure $f \in \mathcal{P}(\mathcal{C}([0,T];\mathbb{R}^d) \times \mathcal{C}([0,T];\mathbb{R}^d) \times \mathcal{C}([0,T];\mathbb{R}^d))$, which

is the unique weak solution to the following mean-field PSO system with local best characterized by the nonlinear Vlasov-Fokker-Planck equation

$$\partial_{t}f + v \cdot \nabla_{x}f + \nabla_{y} \cdot \left(\nu(x - y)S^{\beta}(x, y)f\right) =$$

$$\nabla_{v} \cdot \left(\frac{\gamma}{m}vf + \frac{\lambda_{1}}{m}(x - y)f + \frac{\lambda_{2}}{m}(x - Y^{\alpha}(\overline{\rho}))f\right)$$

$$+ \left(\frac{\sigma_{2}^{2}}{2m^{2}}D(x - Y^{\alpha}(\overline{\rho}))^{2} + \frac{\sigma_{1}^{2}}{2m^{2}}D(x - y)^{2}\right)\nabla_{v}f\right),$$
(3.36)

where $\overline{\rho}(t,y) = \int_{\mathbb{R}^{2d}} f(t,dx,y,dv)$.

This can be summarized in the following theorem

Theorem 2: Let \mathcal{F} satisfy Assumption 1 and $f_0 \in \mathcal{P}_4(\mathbb{R}^{3d})$. For any $N \geq 2$, we assume that $\{(X_t^{i,N}, Y_t^{i,N}, V_t^{i,N})_{t \in [0,T]}\}_{i=1}^N$ is the unique solution to the SD-PSO system (3.33) with $f_0^{\otimes N}$ -distributed initial data $\{X_0^{i,N}, Y_0^{i,N}, V_0^{i,N}\}_{i=1}^N$. Then the limit (denoted by f) of the sequence of the empirical measure $f^N = \frac{1}{N} \sum_{i=1}^N \delta_{(X_i^{i,N}, Y_i^{i,N}, V_i^{i,N})}$ exists. Moreover, f is the unique weak solution to MF-PSO equation (3.36).

4. Zero-inertia limit and consensus-based optimization

In this section we consider the asymptotic behavior of the previous Vlasov-Fokker-Planck equations modelling the PSO dynamic in the small inertia limit, i.e. $m \to 0$. We will derive the corresponding macroscopic equations which permit to recover and generalize the recently introduced consensus-based optimization (CBO) methods [52]. We refer to [74,77] for a theoretical background concerning the related problem of the overdamped limit of nonlinear Vlasov-Fokker-Planck systems.

4.1. The case without memory effects

Let us first consider the simplified setting in absence of local best. Now we write down the so called McKean-Vlasov process [86] underlying PSO equation (3.4), which is of the form^b

$$\begin{split} &d\overline{X}_{t}^{m} = \overline{V}_{t}^{m}dt\,,\\ &d\overline{V}_{t}^{m} = -\frac{\gamma}{m}\overline{V}_{t}^{m}dt + \frac{\lambda}{m}(X^{\alpha}(\rho_{t}^{m}) - \overline{X}_{t}^{m})dt \end{split} \tag{4.1a}$$

^bWe used the superscript m to emphasize its dependence on the inertia coefficient m.

$$+\frac{\sigma}{m}D(X^{\alpha}(\rho_t^m)-\overline{X}_t^m)dB_t, \qquad (4.1b)$$

where

$$X^{\alpha}(\rho_t^m) = \frac{\int_{\mathbb{R}^d} x \omega_{\alpha}^{\mathcal{E}}(x) \rho^m(t, dx)}{\int_{\mathbb{R}^d} \omega_{\alpha}^{\mathcal{E}}(x) \rho^m(t, dx)}, \quad \rho^m(t, x) = \int_{\mathbb{R}^d} f^m(t, x, dv), \quad (4.2)$$

and the initial data $(\overline{X}_0, \overline{V}_0)$ is the same as in (3.1). Here $f^m(t,x,v)$ is the distribution of $(\overline{X}_t^m, \overline{V}_t^m)$ at time t, which makes the set of equations (4.1) nonlinear. A direct application of the Itô-Doeblin formula yields that the law $f_t^m := f^m(t,\cdot,\cdot)$ at time t is a weak solution to (3.4).

To illustrate the limiting procedure, let us observe that for $m \to 0^+$ from the equation (4.1b) we formally have

$$\overline{V}_t^0 dt = \lambda \left(X^{\alpha}(\rho_t^0) - \overline{X}_t^0 \right) dt + \sigma D(X^{\alpha}(\rho_t^0) - \overline{X}_t^0) dB_t,$$

where we used the fact that $\gamma = 1 - m \to 1$. Substituting the above identity into the equation (4.1a) and omitting the superscripts gives the first order CBO system [52]

$$d\overline{X}_t = \lambda (X^{\alpha}(\rho_t) - \overline{X}_t)dt + \sigma D(X^{\alpha}(\rho_t) - \overline{X}_t)dB_t.$$
 (4.3)

Therefore, the CBO models based on a multiplicative noise can be understood as reduced order approximations of SD-PSO dynamics.

4.1.1. Formal derivation in the mean-field case

In the sequel we will develop these arguments in the case of the nonlinear Vlasov-Fokker-Planck equation (3.4) describing the evolution of the distribution of (4.1). We re-write the scaled Vlasov-Fokker-Planck system in the form

$$\partial_t f + v \cdot \nabla_x f + \frac{1}{m} \nabla_v \cdot (mvf + \lambda(X^{\alpha}(\rho) - x)f) = L_m(f)$$
 (4.4)

where we used the fact that $\gamma = 1 - m$ and define

$$L_m(f) = \frac{1}{m} \nabla_v \cdot \left(vf + \frac{\sigma^2}{2m} D(x - X^{\alpha}(\rho))^2 \nabla_v f \right)$$
$$= \frac{1}{m} \sum_{j=1}^d \frac{\sigma^2}{2} (x_j - X_j^{\alpha}(\rho))^2 \frac{\partial}{\partial v_j} \left(\frac{2fv_j}{\sigma^2 (x_j - X_j^{\alpha}(\rho))^2} + \frac{1}{m} \frac{\partial f}{\partial v_j} \right).$$

Note that the last equality is a direct consequence of identity (3.5). Let us now introduce the local Maxwellian with unitary mass and zero momentum

$$\mathcal{M}_{m}(x, v, t) = \prod_{j=1}^{d} M_{m}(x_{j}, v_{j}, t),$$

$$M_{m}(x_{j}, v_{j}, t) = \frac{m^{1/2}}{\pi^{1/2} \sigma |x_{j} - X_{i}^{\alpha}(\rho)|} \exp \left\{ -\frac{m v_{j}^{2}}{\sigma^{2} (x_{j} - X_{i}^{\alpha}(\rho))^{2}} \right\},$$

then we have

$$L_m(f) = \frac{1}{m^2} \sum_{j=1}^d \frac{\sigma^2}{2} (x_j - X_j^{\alpha}(\rho))^2 \frac{\partial}{\partial v_j} \left(f \frac{\partial}{\partial v_j} \log \left(\frac{f}{M_m(x_j, v_j, t)} \right) \right).$$

Therefore $L_m(f)$ is of order $1/m^2$ and we can write for small values of $m \ll 1$

$$f(x, v, t) = \rho(x, t)\mathcal{M}_m(x, v, t). \tag{4.5}$$

Let us now integrate equation (4.4) with respect to v, and multiply the same equation by v and ingrate again with respect to v, we get

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) = 0$$

$$\frac{\partial \rho u}{\partial t} + \int_{\mathbb{R}^d} v \left(v \cdot \nabla_x f \right) dv = -\frac{\gamma}{m} \rho u + \frac{1}{m} \lambda (X^{\alpha}(\rho) - x) \rho$$

where

$$\rho u = \int_{\mathbb{D}^d} f(x, v, t) v \, dv.$$

Now assuming (4.5) we can compute for $m \ll 1$ the j-th component of the second term in the right hand side of last equation as

$$\int_{\mathbb{R}^d} v_j \left(v \cdot \nabla_x \left(\rho(x, t) \mathcal{M}_m(x, v, t) \right) \right) dv = \sum_{j=1}^d \frac{\partial}{\partial x_j} \left(\rho(x, t) \int_{\mathbb{R}^d} v_j (v_j \mathcal{M}_m(x, v, t)) dv \right)$$

$$= \frac{\partial}{\partial x_j} \left(\rho(x, t) \int_{\mathbb{R}} v_j^2 M_m(x_j, v_j, t) dv_j \right)$$

$$= \frac{\sigma^2}{2m} \frac{\partial}{\partial x_j} \left(\rho(x, t) (x_j - X_j^{\alpha}(\rho))^2 \right)$$

which provides the macroscopic PSO system without local best

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) = 0,$$

$$\frac{\partial (\rho u)_j}{\partial t} + \frac{\sigma^2}{2m} \frac{\partial}{\partial x_j} \left(\rho(x, t) (x_j - X_j^{\alpha}(\rho))^2 \right) =$$

$$= -\frac{1 - m}{m} (\rho u)_j + \frac{1}{m} \lambda (X_j^{\alpha}(\rho) - x_j) \rho.$$
(4.6)

Formally, as $m \to 0^+$, from the second equation in (4.6) we get

$$(\rho u)_j = \lambda (X_j^{\alpha}(\rho) - x_j)\rho - \frac{\sigma^2}{2} \frac{\partial}{\partial x_j} \left(\rho(x, t) (x_j - X_j^{\alpha}(\rho))^2 \right),$$

which substituted in the first equation yields the mean-field CBO system [52]

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot \lambda (X^{\alpha}(\rho) - x)\rho = \frac{\sigma^2}{2} \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} \left(\rho(x, t) (x_j - X_j^{\alpha}(\rho))^2 \right). \tag{4.7}$$

Therefore, in the small inertia limit we expect the macroscopic density in the PSO system (3.4) to be well approximated by the solution of the CBO equation (4.7). We emphasize that system (4.6) represents a novel mean-field optimization model with an intermediate level of description between the mean-field PSO system (4.4) and the mean-field CBO system (4.7).

4.1.2. Rigorous derivation

In this section, we present a rigorous derivation of the zero-inertia limit [71]. More precisely we prove that as $m \to 0^+$, the processes $\{\overline{X}^m\}$ satisfying the SDEs (4.1) converge weakly to the solution \overline{X} to the SDE (4.3) in the continuous path space $\mathcal{C}([0,T];\mathbb{R}^d)$, and a convergence rate is obtained. The main theorem can be stated as below:

Theorem 3: Let Assumption 1 hold and $(X_t^m, V_t^m)_{t \in [0,T]}$ satisfy the system (4.1). Then as $m \to 0^+$, the sequence of stochastic processes $\{\overline{X}^m\}_{0 < m \leq \frac{1}{2}}$ converge weakly to \overline{X} , which is the unique solution to the

following SDE:

$$\overline{X}_{t} = \overline{X}_{0} + \lambda \int_{0}^{t} (X^{\alpha}(\rho_{s}) - \overline{X}_{s}) ds + \sigma \int_{0}^{t} D(X^{\alpha}(\rho_{s}) - \overline{X}_{s}) dB_{s}.$$

$$(4.8)$$

Moreover it holds that

$$\sup_{t \in [0,T]} \mathbb{E}[|\overline{X}_t^m - \overline{X}_t|^2] \le C m, \qquad (4.9)$$

where the constant C depends only on $\mathbb{E}[|\overline{X}_0|^4]$, $\mathbb{E}[|\overline{V}_0|^4]$, M, C_u , C_l , λ , σ , d, and T.

Remark 4.1: It follows from the definition of Wasserstein distance that

$$\sup_{t \in [0,T]} W_2^2(\rho_t^m, \rho_t) \le \sup_{t \in [0,T]} \mathbb{E}[|\overline{X}_t^m - \overline{X}_t|^2] \le C m, \qquad (4.10)$$

which in a way is consistent with the result obtained in [77, Theorem 1.3], where the authors obtained a quantified overdamped limit (with the same rate m) of the singular Vlasov-Poisson-Fokker-Planck system to the aggregation-diffusion equation.

The following theorem gives the well-posedness of the mean-field PSO dynamic (4.1) whose proof is analogous to [69, Theorem 2.3] or [51, Theorem 3.1], and thus omitted.

Theorem 4.2: Let Assumption 1 hold. If $(\overline{X}_0^m, \overline{V}_0^m) = (\overline{X}_0, \overline{V}_0)$ is distributed according to f_0 with $f_0 \in \mathcal{P}_4(\mathbb{R}^{2d})$, then for each T > 0 and $m \in (0,1]$, the nonlinear SDE (4.1) admits a unique strong solution up to time T with the initial data $(\overline{X}_0^m, \overline{V}_0^m)$ and it holds further that

$$\sup_{t \in [0,T]} \mathbb{E}\left[|\overline{X}_t^m|^4 + |\overline{V}_t^m|^4 \right] \le e^{CT} \cdot \mathbb{E}\left[|\overline{X}_0|^4 + |\overline{V}_0|^4 \right], \tag{4.11}$$

where C depends only on $\lambda, m, \sigma, M, C_u$, and C_l .

Solving (4.1b) for \overline{V}_t^m gives

$$\overline{V}_{t}^{m} = e^{-\frac{\gamma}{m}t}\overline{V}_{0} + \frac{\lambda}{m} \int_{0}^{t} e^{-\frac{\gamma}{m}(t-s)} (X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) ds + \frac{\sigma}{m} \int_{0}^{t} e^{-\frac{\gamma}{m}(t-s)} D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s},$$

which implies that

$$\overline{X}_{t}^{m} = \overline{X}_{0} + \int_{0}^{t} \overline{V}_{\tau} d\tau = \overline{X}_{0} + \int_{0}^{t} e^{-\frac{\gamma}{m}\tau} \overline{V}_{0} d\tau
+ \frac{\lambda}{m} \int_{0}^{t} \int_{0}^{\tau} e^{-\frac{\gamma}{m}(\tau - s)} (X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) ds d\tau
+ \frac{\sigma}{m} \int_{0}^{t} \int_{0}^{\tau} e^{-\frac{\gamma}{m}(\tau - s)} D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s} d\tau .$$
(4.12)

Then \overline{X}_t^m has the law ρ_t^m for each $t \geq 0$. Each continuous stochastic process \overline{X}^m may be seen as a $\mathcal{C}([0,T];\mathbb{R}^d)$ valued random function and it induces a probability measure (or law, denoted by ρ^m) on $\mathcal{C}([0,T];\mathbb{R}^d)$. We shall use the weak convergence in the space of probability measures on $\mathcal{C}([0,T];\mathbb{R}^d)$. In what follows, we write $\overline{X}^m \rightharpoonup \overline{X}$ or $\rho^m \rightharpoonup \rho$ with ρ being the law of \overline{X} , if $\{\rho^m\}_{m>0}$, as a sequence of probability measures, converges weakly to ρ , i.e., for each bounded continuous functional Φ on $\mathcal{C}([0,T];\mathbb{R}^d)$, there holds $\lim_{m\to 0^+} \mathbb{E}\left[\Phi(\overline{X}^m)\right] =$ $\mathbb{E}\left[\Phi(\overline{X})\right]$. The weak convergence $\overline{X}^m \rightharpoonup \overline{X}$ is stronger than and actually implies the convergence of $\{\rho_t^m\}_{m>0}$ to ρ_t with ρ_t being the law of \overline{X}_t for each $t \geq 0$, while the converse need not hold. Moreover, due to the separability and completeness of the space $\mathcal{C}([0,T];\mathbb{R}^d)$, Prohorov's theorem implies that the relative compactness is equivalent to the tightness; see [85] for more details.

Theorem 4.3: Let Assumption 1 hold and $(X_t^m, V_t^m)_{t \in [0,T]}$ satisfy the system (4.1). For each countable subsequence $\{m_k\}_{k\in\mathbb{N}}\subset [0,\frac{1}{2}]$ with $\lim_{k\to\infty} m_k = 0$, the sequence of probability distributions $\{\rho^{m_k}\}_{k\in\mathbb{N}}$ of $\{\overline{X}^{m_k}\}_{k\in\mathbb{N}}$ is tight.

Proof: By Lemma 3.2, it is sufficient to justify conditions (Con1) and (Con2) in Aldous tightness criteria.

• Step 1: Checking (Con1). First, for $0 < m \le \frac{1}{2}$, recalling (4.12), we have by Fubini's theorem (see [87, Theorem 4.33] for the stochastic version)

$$\begin{split} \overline{X}_t^m &= \overline{X}_0 + \int_0^t e^{-\frac{\gamma}{m}\tau} \overline{V}_0 d\tau + \frac{\lambda}{m} \int_0^t \int_0^\tau e^{-\frac{\gamma}{m}(\tau - s)} (X^\alpha(\rho_s^m) - \overline{X}_s^m) ds d\tau \\ &+ \frac{\sigma}{m} \int_0^t \int_0^\tau e^{-\frac{\gamma}{m}(\tau - s)} D(X^\alpha(\rho_s^m) - \overline{X}_s^m) dB_s d\tau \\ &= \overline{X}_0 + \int_0^t e^{-\frac{\gamma}{m}\tau} \overline{V}_0 d\tau + \frac{\lambda}{m} \int_0^t \int_s^t e^{-\frac{\gamma}{m}(\tau - s)} d\tau (X^\alpha(\rho_s^m) - \overline{X}_s^m) ds \end{split}$$

$$+\frac{\sigma}{m}\int_{0}^{t}\int_{s}^{t}e^{-\frac{\gamma}{m}(\tau-s)}d\tau D(X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m})dB_{s}$$

$$=\overline{X}_{0}+\frac{m}{\gamma}(1-e^{-\frac{\gamma}{m}t})\overline{V}_{0}+\frac{\lambda}{\gamma}\int_{0}^{t}(1-e^{-\frac{\gamma}{m}(t-s)})(X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m})ds$$

$$+\frac{\sigma}{\gamma}\int_{0}^{t}(1-e^{-\frac{\gamma}{m}(t-s)})D(X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m})dB_{s}.$$

$$(4.13)$$

Note here the assumption on $0 < m \le \frac{1}{2}$ ensures that $\gamma = 1 - m \in [\frac{1}{2}, 1)$, so $\frac{1}{\gamma}$ is well defined. It follows from Hölder's inequality that

$$|\overline{X}_{t}^{m}|^{4} \leq 64|\overline{X}_{0}|^{4} + \frac{64m^{4}}{\gamma^{4}}|\overline{V}_{0}|^{4} + \frac{64\lambda^{4}t^{3}}{\gamma^{4}} \int_{0}^{t} |X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}|^{4}ds + \frac{64\sigma^{4}}{\gamma^{4}} \left| \int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s} \right|^{4}, \quad (4.14)$$

where we have used the fact that for any sequence $\{a_i\}_{i=1}^n \geq 0$ and $p \geq 2$, there holds

$$\left(\sum_{i=1}^n a_i\right)^p \le n^{p-1} \sum_{i=1}^n a_i^p.$$

Using the moment inequality for stochastic integrals as in [88, Theorem 7.1] yields that

$$\mathbb{E}\left[\left|\int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s}\right|^{4}\right]$$

$$\leq d^{3}\mathbb{E}\left[\sum_{k=1}^{d} \left|\int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) (X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m})_{k} dB_{s}^{k} e_{k}\right|^{4}\right]$$

$$\leq 36d^{3}t \int_{0}^{t} \mathbb{E}\left[\sum_{k=1}^{d} |(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m})_{k}|^{4}\right] ds$$

$$\leq 36d^{3}t \int_{0}^{t} \mathbb{E}\left[|X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}|^{4}\right] ds.$$

Thus,

$$\begin{split} \mathbb{E}[|\overline{X}_{t}^{m}|^{4}] &\leq 64 \mathbb{E}[|\overline{X}_{0}|^{4}] + \frac{64 m^{4}}{\gamma^{4}} \mathbb{E}[|\overline{V}_{0}|^{4}] \\ &+ \frac{64 (\lambda^{4} t^{3} + 36 d^{3} t \sigma^{4})}{\gamma^{4}} \int_{0}^{t} \mathbb{E}[|X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}|^{4}] ds \,. \end{split}$$

Notice that

$$\mathbb{E}[|X^{\alpha}(\rho_{t}^{m}) - \overline{X}_{t}^{m}|^{4}] \leq 8|X^{\alpha}(\rho_{t}^{m})|^{4} + 8\mathbb{E}[|\overline{X}_{t}^{m}|^{4}]$$

$$\leq 8(b_{1} + b_{2}\mathbb{E}[|\overline{X}_{t}^{m}|^{2}])^{2} + 8\mathbb{E}[|\overline{X}_{t}^{m}|^{4}]$$

$$\leq c_{1} + c_{2}\mathbb{E}[|\overline{X}_{t}^{m}|^{4}],$$
(4.15)

where we have used Lemma 3.5 in the second inequality, and c_1, c_2 depend only on C_u , M and C_l . Thus we have

$$\mathbb{E}[|\overline{X}_{t}^{m}|^{4}] \leq 64\mathbb{E}[|\overline{X}_{0}|^{4}] + \frac{64m^{4}}{\gamma^{4}}\mathbb{E}[|\overline{V}_{0}|^{4}] + c_{3} + \frac{64c_{2}(\lambda^{4}t^{3} + 36d^{3}t\sigma^{4})}{\gamma^{4}} \int_{0}^{t} \mathbb{E}[|\overline{X}_{s}^{m}|^{4}]ds.$$

Using Gronwall's inequality leads to

$$\mathbb{E}[|\overline{X}_t^m|^4] \le \left(64\mathbb{E}[|\overline{X}_0|^4] + \frac{64m^4}{\gamma^4}\mathbb{E}[|\overline{V}_0|^4] + c_3\right) \cdot \exp\left(\frac{64c_2(\lambda^4 T^3 + 36d^3 T\sigma^4)}{\gamma^4}T\right),\tag{4.16}$$

for all $t \in [0,T]$. Recalling $0 \le m \le \frac{1}{2}$ and $\frac{1}{\gamma} = \frac{1}{1-m} \le 2$, from estimate (4.16) we obtain the boundedness:

$$\mathbb{E}[|\overline{X}_t^m|^4] \le C(\mathbb{E}[|\overline{X}_0|^4], \mathbb{E}[|\overline{V}_0|^4], M, C_u, C_l, \lambda, d, \sigma, T). \tag{4.17}$$

This yields that

$$\sup_{m \in (0,1]} \sup_{t \in [0,T]} \mathbb{E}[|\overline{X}_t^m|^4]$$

$$\leq C(\mathbb{E}[|\overline{X}_0|^4], \mathbb{E}[|\overline{V}_0|^4], M, C_u, C_l, \lambda, \sigma, d, T) =: C_1$$

$$(4.18)$$

where the constant $C_1 > 0$ is independent of m. Therefore, for any $\varepsilon > 0$, there exists a compact subset $K_{\varepsilon} := \{x : |x|^4 \leq \frac{C_1}{\varepsilon}\}$ such that by Markov's inequality

$$\rho_t^m((K_{\varepsilon})^c) = \mathbb{P}(|X_t^m|^4 > \frac{C_1}{\varepsilon}) \le \frac{\varepsilon \mathbb{E}[|X_t^m|^4]}{C_1} \le \varepsilon, \quad \forall \ 0 < m \le 1. \quad (4.19)$$

This means that for each $t \in [0, T]$, each countable subset of $\{\rho_t^m\}_{0 < m \le 1}$ is tight, which verifies condition (Con1) in Lemma 3.2.

• Step 2: Checking (Con2). Let β be a $\sigma(X_s^m; s \in [0, T])$ -stopping time with discrete values such that $\beta + \delta_0 \leq T$. Without any loss of generality, we may assume that the concerned countable subsequence $\{m_k\}_{k\in\mathbb{N}}\subset[0, 1]$

satisfies $m_k \leq \frac{1}{2}$ for all $k \in \mathbb{N}$; thus, we may just consider the case of $0 < m \leq \frac{1}{2}$ which indicates $\frac{1}{2} \leq \gamma < 1$. Recall (4.12) and compute

$$\begin{split} \overline{X}_{\beta+\delta}^{m} - \overline{X}_{\beta}^{m} &= \int_{\beta}^{\beta+\delta} \overline{V}_{\tau} d\tau = \int_{\beta}^{\beta+\delta} e^{-\frac{\gamma}{m}\tau} \overline{V}_{0} d\tau \\ &+ \frac{\lambda}{m} \int_{\beta}^{\beta+\delta} \int_{0}^{\tau} e^{-\frac{\gamma}{m}(\tau-s)} (X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) ds d\tau \\ &+ \frac{\sigma}{m} \int_{\beta}^{\beta+\delta} \int_{0}^{\tau} e^{-\frac{\gamma}{m}(\tau-s)} D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s} d\tau \end{split}$$

$$= \int_{\beta}^{\beta+\delta} e^{-\frac{\gamma}{m}\tau} \overline{V}_{0} d\tau \\ &+ \frac{\lambda}{m} \int_{0}^{\beta} \int_{\beta}^{\beta+\delta} e^{-\frac{\gamma}{m}(\tau-s)} d\tau (X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) ds \\ &+ \frac{\lambda}{m} \int_{\beta}^{\beta+\delta} \int_{s}^{\beta+\delta} e^{-\frac{\gamma}{m}(\tau-s)} d\tau (X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) ds \\ &+ \frac{\sigma}{m} \int_{0}^{\beta} \int_{\beta}^{\beta+\delta} e^{-\frac{\gamma}{m}(\tau-s)} d\tau D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s} \\ &+ \frac{\sigma}{m} \int_{\epsilon}^{\beta+\delta} \int_{\epsilon}^{\beta+\delta} e^{-\frac{\gamma}{m}(\tau-s)} d\tau D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s} \,. \end{split}$$

Then it yields

$$\overline{X}_{\beta+\delta}^{m} - \overline{X}_{\beta}^{m} = \frac{m}{\gamma} \left(e^{-\frac{\gamma}{m}\beta} - e^{-\frac{\gamma}{m}(\beta+\delta)} \right) \overline{V}_{0}
+ \frac{\lambda}{\gamma} \int_{0}^{\beta} \left(e^{-\frac{\gamma}{m}(\beta-s)} - e^{-\frac{\gamma}{m}(\beta+\delta-s)} \right) \left(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m} \right) ds
+ \frac{\lambda}{\gamma} \int_{\beta}^{\beta+\delta} \left(1 - e^{-\frac{\gamma}{m}(\beta+\delta-s)} \right) \left(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m} \right) ds
+ \frac{\sigma}{\gamma} \int_{0}^{\beta} \left(e^{-\frac{\gamma}{m}(\beta-s)} - e^{-\frac{\gamma}{m}(\beta+\delta-s)} \right) D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s}
+ \frac{\sigma}{\gamma} \int_{c}^{\beta+\delta} \left(1 - e^{-\frac{\gamma}{m}(\beta+\delta-s)} \right) D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s}.$$
(4.20)

Note that there holds $|e^{-x} - e^{-y}| \le |x - y| \land 1$ for all $x, y \in [0, \infty)$. Basic computations further indicate that for each $q \ge 1$ and $\tau \in [0, T]$,

$$\int_0^\tau \left| e^{-\frac{\gamma(\tau-s)}{m}} - e^{-\frac{\gamma(\tau+\delta-s)}{m}} \right|^q ds \le \int_0^\tau \left(e^{-\frac{\gamma(\tau-s)}{m}} - e^{-\frac{\gamma(\tau+\delta-s)}{m}} \right) ds$$

$$\begin{split} &= \frac{m}{\gamma} \left(1 - e^{-\frac{\gamma \delta}{m}} \right) - \frac{m}{\gamma} \left(e^{-\frac{\gamma \tau}{m}} - e^{-\frac{\gamma (\tau + \delta)}{m}} \right) \\ &\leq \frac{m}{\gamma} \cdot \frac{\gamma \delta}{m} = \delta, \end{split}$$

and in particular,

$$\int_{\beta}^{\beta+\delta} \left(1 - e^{-\frac{\gamma(\beta+\delta-s)}{m}}\right)^q ds \le \int_{\beta}^{\beta+\delta} 1 \, ds = \delta.$$

Then, it is obvious that

$$\mathbb{E}\left[\left|\frac{m}{\gamma}(e^{-\frac{\gamma}{m}\beta}-e^{-\frac{\gamma}{m}(\beta+\delta)})\overline{V}_0\right|^2\right] \leq \frac{m^2}{\gamma^2} \cdot \frac{\gamma^2\delta^2}{m^2} \left(\mathbb{E}[|\overline{V}_0|^4]\right)^{\frac{1}{2}} \leq \delta^2 \left(\mathbb{E}[|\overline{V}_0|^4]\right)^{\frac{1}{2}}.$$

Next, it follows that

$$\begin{split} & \mathbb{E}\left[\left|\int_{0}^{\beta}(e^{-\frac{\gamma}{m}(\beta-s)}-e^{-\frac{\gamma}{m}(\beta+\delta-s)})(X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m})ds\right|^{2}\right] \\ & \leq \mathbb{E}\left[\int_{0}^{\beta}|e^{-\frac{\gamma}{m}(\beta-s)}-e^{-\frac{\gamma}{m}(\beta+\delta-s)}|^{2}ds\cdot\int_{0}^{\beta}|X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m}|^{2}ds\right] \\ & \leq \delta\cdot T\sup_{s\in[0,T]}\left(\mathbb{E}\left[|X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m}|^{4}\right]\right)^{1/2}, \end{split}$$

and analogously,

$$\begin{split} & \mathbb{E}\left[\left|\int_{\beta}^{\beta+\delta} (1-e^{-\frac{\gamma}{m}(\beta+\delta-s)})(X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m})ds\right|^{2}\right] \\ & \leq \mathbb{E}\left[\int_{\beta}^{\beta+\delta} \left(1-e^{-\frac{\gamma(\beta+\delta-s)}{m}}\right)^{2}ds \cdot \int_{\beta}^{\beta+\delta} |X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m}|^{2}ds\right] \\ & \leq \delta \cdot \mathbb{E}\left[\int_{\beta}^{\beta+\delta} |X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m}|^{2}ds\right] \\ & \leq \delta \cdot T \sup_{s \in [0,T]} \left(\mathbb{E}\left[|X^{\alpha}(\rho_{s}^{m})-\overline{X}_{s}^{m}|^{4}\right]\right)^{1/2}. \end{split}$$

Further, applying Itô's isometry gives

$$\mathbb{E}\left[\left|\int_{0}^{\beta} (e^{-\frac{\gamma}{m}(\beta-s)} - e^{-\frac{\gamma}{m}(\beta+\delta-s)})D(X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m})dB_{s}\right|^{2}\right]$$

$$\leq d\mathbb{E}\left[\int_{0}^{\beta} |e^{-\frac{\gamma}{m}(\beta-s)} - e^{-\frac{\gamma}{m}(\beta+\delta-s)}|^{2}|X^{\alpha}(\rho_{s}^{m}) - \overline{X}_{s}^{m}|^{2}ds\right]$$

$$\leq d \left(\mathbb{E} \left[\int_0^\beta |e^{-\frac{\gamma}{m}(\beta - s)} - e^{-\frac{\gamma}{m}(\beta + \delta - s)}|^4 ds \right] \right)^{1/2} \cdot \left(\mathbb{E} \left[\int_0^\beta |X^\alpha(\rho_s^m) - \overline{X}_s^m|^4 ds \right] \right)^{1/2} \\ \leq d\delta^{1/2} \left(T \sup_{s \in [0, T]} \mathbb{E} \left[|X^\alpha(\rho_s^m) - \overline{X}_s^m|^4 \right] \right)^{1/2},$$

and analogously,

$$\mathbb{E}\left[\left|\int_{\beta}^{\beta+\delta} (1 - e^{-\frac{\gamma}{m}(\beta+\delta-s)}) D(X^{\alpha}(\rho_s^m) - \overline{X}_s^m) dB_s\right|^2\right]$$

$$\leq d\delta^{1/2} \left(T \sup_{s \in [0,T]} \mathbb{E}\left[|X^{\alpha}(\rho_s^m) - \overline{X}_s^m|^4\right]\right)^{1/2}.$$

Therefore, summing up the above estimates and recalling $0 < m \le m_0 = \frac{1}{2}$, $\frac{1}{\gamma} \le 2$, and the relations (4.15) and (4.18), we arrive at

$$\begin{split} & \mathbb{E}[|\overline{X}_{\beta+\delta}^m - \overline{X}_{\beta}^m|^2] \\ & \leq \frac{5}{\gamma^2} \delta^2 (\mathbb{E}[|\overline{V}_0|^4])^{\frac{1}{2}} + \frac{10}{\gamma^2} \left(\lambda^2 \delta T + \sigma^2 d \left(\delta T\right)^{1/2}\right) \sup_{s \in [0,T]} \left(\mathbb{E}\left[|X^{\alpha}(\rho_s^m) - \overline{X}_s^m|^4\right]\right)^{1/2} \\ & \leq C \left(\mathbb{E}[|\overline{X}_0|^4], \mathbb{E}[|\overline{V}_0|^4], M, C_u, C_l, \lambda, \sigma, d, T\right) \left(\delta^{\frac{1}{2}} + \delta + \delta^2\right). \end{split}$$

Hence, for any $\varepsilon > 0$, $\eta > 0$, there exists some $\delta_0 > 0$ such that for all $0 < m \le \frac{1}{2}$ it holds that

$$\sup_{\delta \in [0,\delta_0]} \mathbb{P}(|\overline{X}_{\beta+\delta}^m - \overline{X}_{\beta}^m|^2 \ge \eta) \le \sup_{\delta \in [0,\delta_0]} \frac{\mathbb{E}[|\overline{X}_{\beta+\delta}^m - \overline{X}_{\beta}^m|^2]}{\eta} \le \varepsilon.$$
 (4.21)

This justifies condition Con2 in Lemma 3.2.

Next we shall identify the limit process, before which we recall a lemma on the stability estimate of the nonlinear term $X^{\alpha}(\rho)$.

Lemma 4.4: [51, Lemma 3.2] Assume that $\rho, \widehat{\rho} \in \mathcal{P}_4(\mathbb{R}^d)$. Then the following stability estimate holds

$$|X^{\alpha}(\rho) - X^{\alpha}(\widehat{\rho})| \le CW_2(\rho, \widehat{\rho}), \qquad (4.22)$$

where W_2 is the 2-Wasserstein distance, and C depends only on α, L , $\int_{\mathbb{R}^d} |x|^4 \rho(dx)$, and $\int_{\mathbb{R}^d} |x|^4 \widehat{\rho}(dx)$.

Finally let us prove Theorem 3:

Proof: (Theorem 3) By Theorem 4.3, each subsequence $\{\overline{X}^{m_k}\}_{k\in\mathbb{N}}$ with $m_0 \leq 1/2$ and m_k converging to 0 as $k \to \infty$ admits a subsequence (denoted w.l.o.g. by itself) that converges weakly. By Skorokhod's lemma (see [85, Theorem 6.7 on page 70]) and the existence and uniqueness of strong solution to SDE (4.1), we may find a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the joint processes $\{(\overline{X}^{m_k}, B)\}_{k\in\mathbb{N}}$ converge to some process (\widehat{X}, B) as random variables valued in $\mathcal{C}([0, T]; \mathbb{R}^{2d})$ almost surely. Here B is an identical d-dimensional Wiener process on $(\Omega, \mathcal{F}, \mathbb{P})$. In particular, we have

$$\mathbb{P}\left(\lim_{k\to\infty} \sup_{t\in[0,T]} |\overline{X}_t^{m_k} - \widehat{X}_t| = 0\right) = 1. \tag{4.23}$$

We shall verify that the limit \widehat{X} is indeed the unique solution \overline{X} to SDE (4.8).

Recalling the existence and uniqueness of the strong solution \overline{X}^{m_k} to SDE (4.13) in Theorem 4.2, we have

$$\overline{X}_{t}^{m_{k}} = \overline{X}_{0} + \frac{m_{k}}{\gamma} (1 - e^{-\frac{\gamma}{m_{k}} t}) \overline{V}_{0} + \frac{\lambda}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m_{k}} (t - s)}) (X^{\alpha}(\rho_{s}^{m_{k}}) - \overline{X}_{s}^{m_{k}}) ds
+ \frac{\sigma}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m_{k}} (t - s)}) D(X^{\alpha}(\rho_{s}^{m_{k}}) - \overline{X}_{s}^{m_{k}}) dB_{s}.$$
(4.24)

By the estimates in (4.18) and Fatou's lemma there exists a constant C_2 being independent of m_k such that

$$\sup_{k \in \mathbb{N}} \sup_{t \in [0,T]} \mathbb{E}\left[\left|\overline{X}_{t}^{m_{k}}\right|^{4}\right] + \sup_{t \in [0,T]} \mathbb{E}\left[\left|\widehat{X}_{t}\right|^{4}\right]$$

$$\leq C_{2} := C(\mathbb{E}[\left|\overline{X}_{0}\right|^{4}], \mathbb{E}[\left|\overline{V}_{0}\right|^{4}], C_{\alpha,\mathcal{F}}, \lambda, \sigma, d, T) < \infty. \tag{4.25}$$

As a straightforward consequence of the above boundedness, it holds that

$$\sup_{k \in \mathbb{N}, t \in [0, T]} \mathbb{P}(|\overline{X}_t^{m_k} - \widehat{X}_t| > A) \le \frac{2^4 C_2}{A^4}, \quad \forall A > 0.$$
 (4.26)

Thus, the dominated convergence theorem gives that for each A > 0,

$$\lim_{k \to \infty} \mathbb{E} \left[\int_0^T |\overline{X}_t^{m_k} - \widehat{X}_t|^2 dt \right]$$

$$\leq \limsup_{k \to \infty} \left(\mathbb{E} \left[\int_0^T |\overline{X}_t^{m_k} - \widehat{X}_t|^2 \wedge A^2 dt \right]$$

$$\begin{split} &+ \mathbb{E}\left[\int_{0}^{T} |\overline{X}_{t}^{m_{k}} - \widehat{X}_{t}|^{2} \mathbf{1}_{\{|\overline{X}_{t}^{m_{k}} - \widehat{X}_{t}| > A\}} \, dt\right] \Big) \\ &\leq \limsup_{k \to \infty} \mathbb{E}\left[\int_{0}^{T} |\overline{X}_{t}^{m_{k}} - \widehat{X}_{t}|^{2} \wedge A^{2} \, dt\right] \\ &+ T \cdot \sup_{k \in \mathbb{N}} \sup_{t \in [0,T]} \left(\mathbb{E}\left[|\overline{X}_{t}^{m_{k}} - \widehat{X}_{t}|^{4}\right] \right)^{1/2} \left| \mathbb{P}(|\overline{X}_{t}^{m_{k}} - \widehat{X}_{t}| > A) \right|^{1/2} \\ &\leq \limsup_{k \to \infty} \mathbb{E}\left[\int_{0}^{T} |\overline{X}_{t}^{m_{k}} - \widehat{X}_{t}|^{2} \wedge A^{2} \, dt\right] + \frac{2^{4} \, C_{2} T}{A^{2}} \\ &= \frac{2^{4} \, C_{2} T}{A^{2}}, \end{split}$$

which by the arbitrariness of A > 0 indicates that

$$\lim_{k \to \infty} \mathbb{E} \left[\int_0^T |\overline{X}_t^{m_k} - \widehat{X}_t|^2 dt \right] = 0. \tag{4.27}$$

Letting $\rho(t, dx)$ be the probability distribution of \hat{X}_t for $t \in [0, T]$, Lemma 3.5 gives

$$|X^{\alpha}(\rho_t)| \le (b_1 + b_2 \mathbb{E}[|\widehat{X}_t|^2])^{\frac{1}{2}} \le (b_1 + b_2 C_2^{\frac{1}{2}})^{\frac{1}{2}} =: C_3,$$

and thus

$$\sup_{k \in \mathbb{N}} \sup_{t \in [0,T]} |X^{\alpha}(\rho_t^{m_k})| \le C_3, \quad \text{and} \quad \sup_{t \in [0,T]} |X^{\alpha}(\rho_t)| \le C_3.$$
 (4.28)

Then we compare the SDEs (4.8) and (4.24) term by term. By Lemma 4.4, we have

$$|X^{\alpha}(\rho_t^{m_k}) - X^{\alpha}(\rho_t)|^2 \le CW_2^2(\rho_t^{m_k}, \rho_t) \le C\mathbb{E}[|\overline{X}_t^{m_k} - \widehat{X}_t|^2],$$

and thus by using the fact that $\gamma = 1 - m_k$, one has

$$\mathbb{E}\left[\left|\frac{\lambda}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m_{k}}(t-s)}) (X^{\alpha}(\rho_{s}^{m_{k}}) - \overline{X}_{s}^{m_{k}}) ds - \lambda \int_{0}^{t} (X^{\alpha}(\rho_{s}) - \widehat{X}_{s}) ds\right|^{2}\right] \\
\leq 2\mathbb{E}\left[\left|\frac{\lambda}{1 - m_{k}} \int_{0}^{t} (1 - e^{-\frac{1 - m_{k}}{m_{k}}(t-s)}) \cdot (X^{\alpha}(\rho_{s}^{m_{k}}) - X^{\alpha}(\rho_{s}) + \widehat{X}_{s} - \overline{X}_{s}^{m_{k}}) ds\right|^{2}\right] \\
+ 2\mathbb{E}\left[\left|\lambda \int_{0}^{t} \left(\frac{1 - e^{-\frac{1 - m_{k}}{m_{k}}(t-s)}}{1 - m_{k}} - 1\right) (X^{\alpha}(\rho_{s}) - \widehat{X}_{s}) ds\right|^{2}\right] \\
\leq C\mathbb{E}\left[\int_{0}^{t} \left|\widehat{X}_{s} - \overline{X}_{s}^{m_{k}}\right|^{2} ds\right] + C\lambda^{2} \int_{0}^{t} \left|\frac{1 - e^{-\frac{1 - m_{k}}{m_{k}}(t-s)}}{1 - m_{k}} - 1\right|^{2} ds$$

$$\cdot \mathbb{E}\left[\int_{0}^{T} \left| X^{\alpha}(\rho_{s}) - \widehat{X}_{s} \right|^{2} ds \right] \\
\leq C \mathbb{E}\left[\int_{0}^{t} \left| \widehat{X}_{s} - \overline{X}_{s}^{m_{k}} \right|^{2} ds \right] + C \int_{0}^{t} \left| \frac{1 - e^{-\frac{1 - m_{k}}{m_{k}}(t - s)} - (1 - m_{k})}{1 - m_{k}} \right|^{2} ds \\
\leq C \mathbb{E}\left[\int_{0}^{t} \left| \widehat{X}_{s} - \overline{X}_{s}^{m_{k}} \right|^{2} ds \right] + C \int_{0}^{t} \left(|m_{k}|^{2} + e^{-\frac{2(1 - m_{k})}{m_{k}}(t - s)} \right) ds \\
\leq C \mathbb{E}\left[\int_{0}^{t} \left| \widehat{X}_{s} - \overline{X}_{s}^{m_{k}} \right|^{2} ds \right] + C \left(t \left| m_{k} \right|^{2} + \frac{m_{k}}{2(1 - m_{k})} \right), \tag{4.29}$$

where the constants Cs are independent of k. For the stochastic integrals, it holds analogously that

$$\mathbb{E}\left[\left|\frac{\sigma}{\gamma}\int_{0}^{t}(1-e^{-\frac{\gamma}{m_{k}}(t-s)})D(X^{\alpha}(\rho_{s}^{m_{k}})-\overline{X}_{s}^{m_{k}})dB_{s}\right.\right. \\
\left.\left.-\sigma\int_{0}^{t}D(X^{\alpha}(\rho_{s})-\widehat{X}_{s})dB_{s}\right|^{2}\right] \\
\leq d\sigma^{2}\sum_{n=1}^{d}\mathbb{E}\left[\left|\frac{1}{\gamma}\int_{0}^{t}(1-e^{-\frac{\gamma}{m_{k}}(t-s)})(X^{\alpha}(\rho_{s}^{m_{k}})-\overline{X}_{s}^{m_{k}})_{n}dB_{s}^{n}e_{n}\right. \\
\left.\left.-\int_{0}^{t}(X^{\alpha}(\rho_{s})-\widehat{X}_{s})_{n}dB_{s}^{n}e_{n}\right|^{2}\right] \\
= d\sigma^{2}\sum_{n=1}^{d}\mathbb{E}\left[\int_{0}^{t}\left|\frac{1-e^{-\frac{\gamma}{m_{k}}(t-s)}}{\gamma}(X^{\alpha}(\rho_{s}^{m_{k}})-\overline{X}_{s}^{m_{k}})_{n}\right. \\
\left.-(X^{\alpha}(\rho_{s})-\widehat{X}_{s})_{n}\right|^{2}ds\right]. \tag{4.30}$$

Thus we have

$$\mathbb{E}\left[\left|\frac{\sigma}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m_{k}}(t-s)}) D(X^{\alpha}(\rho_{s}^{m_{k}}) - \overline{X}_{s}^{m_{k}}) dB_{s} - \sigma \int_{0}^{t} D(X^{\alpha}(\rho_{s}) - \widehat{X}_{s}) dB_{s}\right|^{2}\right] \\
\leq 2d\sigma^{2} \sum_{n=1}^{d} \mathbb{E}\left[\int_{0}^{t} \left|\frac{1 - e^{-\frac{\gamma}{m_{k}}(t-s)}}{\gamma} \cdot \left((X^{\alpha}(\rho_{s}^{m_{k}}) - \overline{X}_{s}^{m_{k}})_{n} - (X^{\alpha}(\rho_{s}) - \widehat{X}_{s})_{n}\right)\right|^{2} ds\right] \\
+ 2d\sigma^{2} \sum_{n=1}^{d} \mathbb{E}\left[\int_{0}^{t} \left|\left(\frac{1 - e^{-\frac{\gamma}{m_{k}}(t-s)}}{\gamma} - 1\right)(X^{\alpha}(\rho_{s}) - \widehat{X}_{s})_{n}\right|^{2} ds\right]$$

$$\leq C\mathbb{E}\left[\int_{0}^{t} \left| \overline{X}_{s}^{m_{k}} - \widehat{X}_{s} \right|^{2} ds \right] \\
+ 2d\sigma^{2} \sup_{s \in [0,t]} \mathbb{E}\left[\left| (X^{\alpha}(\rho_{s}) - \widehat{X}_{s}) \right|^{2} \right] \cdot \int_{0}^{t} \left| \left(\frac{1 - e^{-\frac{\gamma}{m_{k}}(t-s)}}{\gamma} - 1 \right) \right|^{2} ds \\
\leq C\mathbb{E}\left[\int_{0}^{t} \left| \widehat{X}_{s} - \overline{X}_{s}^{m_{k}} \right|^{2} ds \right] + C\left(t \left| m_{k} \right|^{2} + \frac{m_{k}}{2(1 - m_{k})} \right). \tag{4.31}$$

In addition, it is obvious that

$$\left| \frac{m_k}{\gamma} (1 - e^{-\frac{\gamma}{m_k} t}) \overline{V}_0 \right| \le C m_k \left| \overline{V}_0 \right|. \tag{4.32}$$

Combining the estimates (4.29)-(4.32), letting k tend to infinity on both sides of (4.24) and recalling $\frac{1}{2} \ge m_k \to 0^+$ and the relation (4.27), we have

$$\widehat{X}_t = \overline{X}_0 + \lambda \int_0^t (X^{\alpha}(\rho_s) - \widehat{X}_s) ds + \sigma \int_0^t D(X^{\alpha}(\rho_s) - \widehat{X}_s) dB_s.$$

Therefore, the limit \widehat{X} turns out to be a solution to SDE (4.8). Meanwhile, in view of the continuity of $X^{\alpha}(\rho)$ in Lemma 4.4, we can easily show that (4.8) admits a unique (strong) solution as in Theorem 4.2 by using Leray-Schauder fixed point theorem as in [51, Theorem 3.1]. Thus, we must have $\widehat{X} = \overline{X}$ that is the unique strong solution to SDE (4.8) with $\sup_{t \in [0,T]} \mathbb{E}\left[|\overline{X}_t|^4\right] \leq C_2$. Further, due to the arbitrariness of the subsequence $\{\overline{X}^{m_k}\}_{k \in \mathbb{N}}$, we conclude that as $m \to 0^+$, the sequence of stochastic processes $\{\overline{X}^m\}_{0 < m \le \frac{1}{2}}$ converge weakly to the unique solution \overline{X} to SDE (4.8).

Finally, to measure the distance between \overline{X}^m and the limit $\widehat{X} = \overline{X}$, we may have similar calculations to (4.29)-(4.32), subtract both sides of SDEs (4.8) from those of (4.24), and arrive at

$$\mathbb{E}[|\overline{X}_t^m - \overline{X}_t|^2] \le C \int_0^t \mathbb{E}[|\overline{X}_s^m - \overline{X}_s|^2] ds + C m, \quad t \in [0, T].$$

By Gronwall's inequality it implies that

$$\sup_{t \in [0,T]} \mathbb{E}[|\overline{X}_t^m - \overline{X}_t|^2] \le Cm, \tag{4.33}$$

where C depends only on $\mathbb{E}[|\overline{X}_0|^4], \mathbb{E}[|\overline{V}_0|^4], C_u, M, C_l, \lambda, \sigma, d$, and T. This completes the proof.

4.2. The general case with memory

Next, we consider the same small inertia scaling in the general case with dependence from the local best. Again, we first write down the nonlinear McKean-Vlasov process corresponding to the SD-PSO system (3.33), which is of the form

$$\begin{split} d\overline{X}_{t}^{m} &= \overline{V}_{t}^{m} dt, \qquad (4.34a) \\ d\overline{Y}_{t}^{m} &= \nu \left(\overline{X}_{t}^{m} - \overline{Y}_{t}^{m} \right) S^{\beta} \left(\overline{X}_{t}^{m}, \overline{Y}_{t}^{m} \right) dt, \qquad (4.34b) \\ d\overline{V}_{t}^{m} &= -\frac{\gamma}{m} \overline{V}_{t}^{m} dt + \frac{\lambda_{1}}{m} \left(\overline{Y}_{t}^{m} - \overline{X}_{t}^{m} \right) dt \\ &+ \frac{\lambda_{2}}{m} \left(Y^{\alpha} (\overline{\rho}_{t}^{m}) - \overline{X}_{t}^{m} \right) dt + \frac{\sigma_{1}}{m} D \left(\overline{Y}_{t}^{m} - \overline{X}_{t}^{m} \right) dB_{t}^{1} \quad (4.34c) \\ &+ \frac{\sigma_{2}}{m} D \left(Y^{\alpha} (\overline{\rho}_{t}^{m}) - \overline{X}_{t}^{m} \right) dB_{t}^{2} \,, \end{split}$$

where B^1 and B^2 are two mutually independent d-dimensional Wiener processes, and similarly to the previous section, we introduce the following regularization of the global best position

$$Y^{\alpha}(\overline{\rho}_{t}^{m}) = \frac{\int_{\mathbb{R}^{d}} y \omega_{\alpha}(y) \overline{\rho}^{m}(t, dy)}{\int_{\mathbb{R}^{d}} \omega_{\alpha}(y) \overline{\rho}^{m}(t, dy)}, \qquad \overline{\rho}^{m}(t, y) = \iint_{\mathbb{R}^{d} \times \mathbb{R}^{d}} f^{m}(t, dx, y, dv).$$

As $m \to 0^+$ we formally get from (4.34c)

$$\overline{V}_{t}^{0}dt = \lambda_{1} \left(\overline{Y}_{t}^{0} - \overline{X}_{t}^{0} \right) dt + \lambda_{2} \left(Y^{\alpha} (\overline{\rho}_{t}^{0}) - \overline{X}_{t}^{0} \right) dt + \sigma_{1} D (\overline{Y}_{t}^{0} - \overline{X}_{t}^{0}) dB_{t}^{1} + \sigma_{2} D (Y^{\alpha} (\overline{\rho}_{t}^{0}) - \overline{X}_{t}^{0}) dB_{t}^{2},$$

which inserted into (4.34a) and omitting the superscripts corresponds to a novel CBO system with local best

$$d\overline{X}_{t} = \lambda_{1}(\overline{Y}_{t} - \overline{X}_{t})dt + \lambda_{2}(Y^{\alpha}(\overline{\rho}_{t}) - \overline{X}_{t})dt + \sigma_{1}D(\overline{Y}_{t} - \overline{X}_{t})dB_{t}^{1} + \sigma_{2}D(Y^{\alpha}(\overline{\rho}_{t}) - \overline{X}_{t})dB_{t}^{2},$$

$$d\overline{Y}_{t} = \nu\left(\overline{X}_{t} - \overline{Y}_{t}\right)S^{\beta}\left(\overline{X}_{t}, \overline{Y}_{t}\right)dt.$$

$$(4.35)$$

In contrast with the model recently introduced in [53] the above first order CBO method avoids backward time integration through the use of an additional differential equation. We refer to [67] for further details on the above CBO system.

4.2.1. Formal derivation in the mean-field case

Concerning the corresponding MF-PSO limit characterized by (3.36) for $m \to 0^+$ we can essentially perform analogous computations as in the previous section (see [68]). Similarly by considering the local Maxwellian with unitary mass and zero momentum

$$\mathcal{M}_m(x, y, v, t) = \prod_{j=1}^d M_m(x_j, y_j, v_j, t),$$

$$M_m(x_j, y_j, v_j, t) = \frac{m^{1/2}}{\pi^{1/2} |\Sigma(x_j, y_j, t)|} \exp\left\{-\frac{mv_j^2}{\Sigma(x_j, y_j, t)^2}\right\},$$

where

$$\Sigma(x_j, y_j, t)^2 = \sigma_2^2 (x_j - Y_j^{\alpha}(\overline{\rho}))^2 + \sigma_1^2 (x_j - y_j)^2,$$

we can assume for $m \ll 1$

$$f(x, y, v, t) = \rho(x, y, t) \mathcal{M}_{\varepsilon}(x, y, v, t). \tag{4.36}$$

After integration of the MF-PSO equation (3.36) with respect to v, we get the second order macroscopic PSO system with local best

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho u) + \nabla_y \cdot (\nu(x - y)S^{\beta}(x, y)\rho) = 0$$

$$\frac{\partial (\rho u)_j}{\partial t} + \frac{\sigma^2}{2m} \frac{\partial}{\partial x_j} (\rho(x, t)\Sigma(x_j, y_j, t)^2) =$$

$$-\frac{\gamma}{m} (\rho u)_j + \frac{1}{m} (\lambda_1 (y_j - x_j) + \lambda_2 (Y_j^{\alpha}(\overline{\rho}) - x_j))\rho.$$
(4.37)

Formally, as $m \to 0^+$, the above system reduces to a novel mean-field CBO system with local best

$$\frac{\partial \rho}{\partial t} + \nabla_x \cdot (\lambda_1 (y - x) + \lambda_2 (Y^{\alpha}(\overline{\rho}) - x)) \rho
+ \nabla_y \cdot (\nu (x - y) S^{\beta}(x, y) \rho)
= \frac{1}{2} \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} \left(\rho(x, t) \left(\sigma_1^2 (x_j - y_j)^2 + \sigma_2^2 (x_j - Y_j^{\alpha}(\overline{\rho}))^2 \right) \right).$$
(4.38)

4.2.2. Rigorous derivation

Since the proof of the zero-inertia limit for the PSO dynamics with memory effects follows similar arguments as developed in section 4.1.2 and no essential innovation is needed to be explained, we only recall the main results here.

Let us solve (4.34c) to obtain

$$\overline{X}_{t}^{m} = \overline{X}_{0} + \frac{m}{\gamma} (1 - e^{-\frac{\gamma}{m}t}) \overline{V}_{0} + \frac{\lambda_{1}}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) (\overline{Y}_{s}^{m} - \overline{X}_{s}^{m}) ds
+ \frac{\sigma_{1}}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) D(\overline{Y}_{s}^{m} - \overline{X}_{s}^{m}) dB_{s}^{1}
+ \frac{\lambda_{2}}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) (Y^{\alpha}(\overline{\rho}_{s}^{m}) - \overline{X}_{s}^{m}) ds
+ \frac{\sigma_{2}}{\gamma} \int_{0}^{t} (1 - e^{-\frac{\gamma}{m}(t-s)}) D(Y^{\alpha}(\overline{\rho}_{s}^{m}) - \overline{X}_{s}^{m}) dB_{s}^{2}$$

$$(4.39)$$

and

$$\overline{Y}_{t}^{m} = \overline{Y}_{0} + \nu \int_{0}^{t} \left(\overline{X}_{s}^{m} - \overline{Y}_{s}^{m} \right) S^{\beta} \left(\overline{X}_{s}^{m}, \overline{Y}_{s}^{m} \right) ds. \tag{4.40}$$

Similar to Theorem 4.3 one can prove the following result of tightness.

Theorem 4.5: Let Assumption 1 hold and $(\overline{X}_t^m, \overline{Y}_t^m, \overline{V}_t^m)_{t \in [0,T]}$ satisfy the system (4.34a)–(4.34c). For each countable subsequence $\{m_k\}_{k \in \mathbb{N}} \subset [0, \frac{1}{2}]$ with $\lim_{k \to \infty} m_k = 0$, the sequence of probability distributions $\{\rho^{m_k}\}_{k \in \mathbb{N}}$ of $\{(\overline{X}^{m_k}, \overline{Y}^{m_k})\}_{k \in \mathbb{N}}$ is tight.

Then following the lines of the proof in Theorem 3, one can obtain

Theorem 4: Let Assumption 1 hold and $(\overline{X}_t^m, \overline{Y}_t^m)_{t \in [0,T]}$ satisfy the system (4.39)–(4.40). Then as $m \to 0^+$, the sequence of stochastic processes $\{(\overline{X}^m, \overline{Y}^m)\}_{0 < m \leq \frac{1}{2}}$ converge weakly to $(\overline{X}, \overline{Y})$ which is the unique solution to the following coupled SDE:

$$\overline{X}_{t} = \overline{X}_{0} + \lambda_{1} \int_{0}^{t} (\overline{Y}_{s} - \overline{X}_{s}) ds + \sigma_{1} \int_{0}^{t} D(\overline{Y}_{s} - \overline{X}_{s}) dB_{s}^{1}$$

$$+ \lambda_{2} \int_{0}^{t} (Y^{\alpha}(\overline{\rho}_{s}) - \overline{X}_{s}) ds + \sigma_{2} \int_{0}^{t} D(Y^{\alpha}(\overline{\rho}_{s}) - \overline{X}_{s}) dB_{s}^{2},$$

$$\overline{Y}_{t} = \overline{Y}_{0} + \nu \int_{0}^{t} (\overline{X}_{s} - \overline{Y}_{s}) S^{\beta} (\overline{X}_{s}, \overline{Y}_{s}) ds.$$

Moreover it holds that

$$\sup_{t \in [0,T]} \mathbb{E}\left[\left|\overline{X}_t^m - \overline{X}_t\right|^2 + \left|\overline{Y}_t^m - \overline{Y}_t\right|^2\right] \le C m, \qquad (4.41)$$

where the constant C depends only on $\mathbb{E}[|\overline{X}_0|^4]$, $\mathbb{E}[|\overline{Y}_0|^4]$, $\mathbb{E}[|\overline{V}_0|^4]$, λ_1 , σ_2 , λ_2 , σ_2 , d, β , T, C_u , M, C_l , and ν .

5. Convergence to the global minimum

In this section we present some results on the global convergence of the PSO model (3.1) without memory effects. The extension to the case with memory effects is not strightforward and is actually under study. Here we will follow the presentation in [72], we refer to [51, 52, 61–63] for similar results for CBO and related models. A different approach to the global convergence of CBO has been presented recently in [58].

Let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be the solution to the nonlinear SDE (4.1) (dropping the superscript m), and consider the quantity

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

then it holds that

$$\mathcal{H}(t) \geq \frac{1}{2} (\frac{\gamma}{2m})^2 |\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + \frac{1}{2} |\overline{V}_t|^2$$

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

$$\leq \frac{3}{2} ((\frac{\gamma}{2m})^2 + 1) (|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2 + |\overline{V}_t|^2).$$
(5.1)

The goal is then to obtain the decay property of $\mathcal{H}(t)$.

In the following we shall use the notation

$$\delta \overline{X}_t := \overline{X}_t - \mathbb{E}[\overline{X}_t], \qquad (5.2)$$

then $\mathbb{E}[|\delta \overline{X}_t|^2]$ is the variance of X_t . Now we can derive an evolution inequality of the quantity $\mathbb{E}[\mathcal{H}(t)]$.

Theorem 5.1: Under the Assumption 1, let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be the solution to the nonlinear SDE (4.1). Then $\mathbb{E}[\mathcal{H}(t)]$ satisfies

$$\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \leq -\frac{\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] \\
-\left(\frac{\lambda\gamma}{2m^2} - (\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2})\frac{2e^{-\alpha\underline{\mathcal{F}}}}{\mathbb{E}[e^{-\alpha\overline{\mathcal{F}}(\overline{X}_t)}]}\right)\mathbb{E}[|\delta\overline{X}_t|^2]. \tag{5.3}$$

Proof: First the integration by parts formula gives

$$\frac{d}{dt}\mathbb{E}[|\delta \overline{X}_t|^2] = 2\mathbb{E}[\delta \overline{X}_t \cdot \overline{V}_t], \qquad (5.4)$$

where we have used the fact that $\mathbb{E}[\delta \overline{X}_t \cdot \mathbb{E}[V_t]] = 0$. Applying Itô-Doeblin formula and taking zero-value of the stochastic integrals, we have for any $\varepsilon > 0$,

$$\frac{d}{dt}\mathbb{E}[|\overline{V}_t|^2] = -2\frac{\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] + 2\frac{\lambda}{m}\mathbb{E}[\overline{V}_t \cdot (X^{\alpha}(\rho_t) - \overline{X}_t)]
+ \frac{\sigma^2}{m^2}\mathbb{E}[|X^{\alpha}(\rho_t) - \overline{X}_t|^2]
\leq -(\frac{2\gamma}{m} - \frac{\lambda}{\varepsilon m})\mathbb{E}[|\overline{V}_t|^2] + (\frac{\varepsilon\lambda}{m} + \frac{\sigma^2}{m^2})\mathbb{E}[|X^{\alpha}(\rho_t) - \overline{X}_t|^2]. (5.5)$$

Further by Itô-Doeblin formula, it holds that

$$\frac{d}{dt}\mathbb{E}[\delta\overline{X}_{t}\cdot\overline{V}_{t}]$$

$$= \mathbb{E}[|\overline{V}_{t}|^{2}] - (\mathbb{E}[\overline{V}_{t}])^{2} - \frac{\gamma}{m}\mathbb{E}[\delta\overline{X}_{t}\cdot\overline{V}_{t}] + \frac{\lambda}{m}\mathbb{E}[\delta\overline{X}_{t}\cdot(X^{\alpha}(\rho_{t}) - \overline{X}_{t})]$$

$$\leq \mathbb{E}[|\overline{V}_{t}|^{2}] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_{t}|^{2}] - \frac{\lambda}{m}\mathbb{E}[|\delta\overline{X}_{t}|^{2}] + \frac{\lambda}{m}\mathbb{E}[\delta\overline{X}_{t}\cdot(X^{\alpha}(\rho_{t}) - \mathbb{E}[\overline{X}_{t}])]$$

$$= \mathbb{E}[|\overline{V}_{t}|^{2}] - \frac{\gamma}{2m}\frac{d}{dt}\mathbb{E}[|\delta\overline{X}_{t}|^{2}] - \frac{\lambda}{m}\mathbb{E}[|\delta\overline{X}_{t}|^{2}].$$
(5.6)

where we have used (5.4) and the fact that $\mathbb{E}[\delta \overline{X}_t \cdot (X^{\alpha}(\rho_t) - \mathbb{E}[\overline{X}_t])] = 0$. Thus, we have

$$\frac{(\gamma)^{2}}{2m} \frac{d}{dt} \mathbb{E}[|\delta \overline{X}_{t}|^{2}] + \frac{\gamma}{2m} \frac{d}{dt} \mathbb{E}[\delta \overline{X}_{t} \cdot \overline{V}_{t}] \\
\leq \frac{\gamma}{2m} \mathbb{E}[|\overline{V}_{t}|^{2}] - \frac{\lambda \gamma}{2m^{2}} \mathbb{E}[|\delta \overline{X}_{t}|^{2}]. \tag{5.7}$$

Collecting estimates (5.5) and (5.7) yields that

$$\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \le -\left(\frac{2\gamma}{m} - \frac{\lambda}{\varepsilon m} - \frac{\gamma}{2m}\right)\mathbb{E}[|\overline{V}_t|^2] - \frac{\lambda\gamma}{2m^2}\mathbb{E}[|\delta\overline{X}_t|^2] + \left(\frac{\varepsilon\lambda}{m} + \frac{\sigma^2}{m^2}\right)\mathbb{E}[|X^{\alpha}(\rho_t) - \overline{X}_t|^2].$$
(5.8)

To estimate the term $\mathbb{E}[|\overline{X}_t - X^{\alpha}(\rho_t)|^2]$, we apply Jensen's inequality to obtain

$$\mathbb{E}[|\overline{X}_t - X^{\alpha}(\rho_t)|^2] \le \frac{\iint |x - y|^2 \omega_{\alpha}^{\mathcal{F}}(y) \rho_t(dy) \rho_t(dx)}{\int \omega_{\alpha}^{\mathcal{F}}(y) \rho_t(dy)}$$

$$\leq 2e^{-\alpha \underline{\mathcal{F}}} \frac{\mathbb{E}[|\delta \overline{X}_t|^2]}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}]}.$$
(5.9)

Hence, by choosing $\varepsilon = \frac{2\lambda}{\gamma}$ we obtain

$$\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \leq -\frac{\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] \\
-\left(\frac{\lambda\gamma}{2m^2} - (\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2})\frac{2e^{-\alpha\underline{\mathcal{F}}}}{\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}]}\right)\mathbb{E}[|\delta\overline{X}_t|^2], \quad (5.10)$$

which completes the proof.

Next we study the evolution of the quantity $\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}]$, and we need an additional assumption on the cost function \mathcal{F} that

A1:
$$\mathcal{F} \in C^2(\mathbb{R}^d)$$
 with $\|\nabla^2 \mathcal{F}\|_{\infty} \leq c_{\mathcal{F}}$ for some constant $c_{\mathcal{F}} > 0$.

Lemma 5.2: Under the Assumption 1 and A1, let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be the solution to the nonlinear SDE (4.1). Then it holds that

$$\frac{d^{2}}{dt^{2}} (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}])^{2} \ge -\frac{\gamma}{m} \frac{d}{dt} (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}])^{2}
-4(\alpha + \frac{\alpha \lambda}{m} 2(\frac{2m}{\gamma})^{2}) c_{\mathcal{F}} e^{-2\alpha \underline{\mathcal{F}}} \mathbb{E}[\mathcal{H}(t)].$$
(5.11)

Proof: First, applying Itô-Doeblin formula and taking zero-value of the stochastic integrals, we have

$$\begin{split} &\frac{d}{dt}\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}] = -\alpha\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}\nabla\mathcal{F}(\overline{X}_t)\cdot\overline{V}_t] \\ &= -\alpha\mathbb{E}[\int_0^t d\langle e^{-\alpha\mathcal{F}(\overline{X}_s)}\nabla\mathcal{F}(\overline{X}_s),\overline{V}_s\rangle] + \alpha\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_0)}\langle\nabla\mathcal{F}(\overline{X}_0),\overline{V}_0\rangle] \\ &= \alpha\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_0)}\langle\nabla\mathcal{F}(\overline{X}_0),\overline{V}_0\rangle] - \alpha\mathbb{E}[\int_0^t \langle e^{-\alpha\mathcal{F}(\overline{X}_s)}\overline{V}_s\nabla^2\mathcal{F}(\overline{X}_s),\overline{V}_s\rangle ds] \\ &+ \alpha^2\mathbb{E}[\int_0^t e^{-\alpha\mathcal{F}(\overline{X}_s)}|\langle\nabla\mathcal{F}(\overline{X}_s),\overline{V}_s\rangle|^2 ds] \\ &- \alpha\mathbb{E}[\int_0^t e^{-\alpha\mathcal{F}(\overline{X}_s)}\langle\nabla\mathcal{F}(\overline{X}_s),-\frac{\gamma}{m}\overline{V}_s\rangle ds] \\ &- \alpha\mathbb{E}[\int_0^t e^{-\alpha\mathcal{F}(\overline{X}_s)}\langle\nabla\mathcal{F}(\overline{X}_s),\frac{\lambda}{m}(X^\alpha(\rho_s)-\overline{X}_s)\rangle ds] \,. \end{split}$$

Further, differentiating both sides with respect to t gives

$$\frac{d^2}{dt^2} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}] = -\alpha \mathbb{E}[\langle e^{-\alpha \mathcal{F}(\overline{X}_t)} \overline{V}_t \nabla^2 \mathcal{F}(\overline{X}_t), \overline{V}_t \rangle]$$

$$+ \alpha^{2} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})} | \langle \nabla \mathcal{F}(\overline{X}_{t}), \overline{V}_{t} \rangle |^{2}]$$

$$- \alpha \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})} \langle \nabla \mathcal{F}(\overline{X}_{t}), -\frac{\gamma}{m} \overline{V}_{t} \rangle]$$

$$- \alpha \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})} \langle \nabla \mathcal{F}(\overline{X}_{t}), \frac{\lambda}{m} (X^{\alpha}(\rho_{t}) - \overline{X}_{t}) \rangle]$$

$$\geq -\frac{\gamma}{m} \frac{d}{dt} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}] - \alpha \mathbb{E}[\langle e^{-\alpha \mathcal{F}(\overline{X}_{t})} \overline{V}_{t} \nabla^{2} \mathcal{F}(\overline{X}_{t}), \overline{V}_{t} \rangle]$$

$$- \alpha \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})} \langle \nabla \mathcal{F}(\overline{X}_{t}), \frac{\lambda}{m} (X^{\alpha}(\rho_{t}) - \overline{X}_{t}) \rangle]$$

$$=: -\frac{\gamma}{m} \frac{d}{dt} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}] + I_{1} + I_{2},$$

$$(5.12)$$

where one has used the fact that

$$-\frac{\gamma}{m}\frac{d}{dt}\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}] = \frac{\alpha\gamma}{m}\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}\langle\nabla\mathcal{F}(\overline{X}_t),\overline{V}_t\rangle]. \tag{5.13}$$

According to assumption A1, it is easy to see that

$$I_1 \ge -\alpha \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)} \|\nabla^2 \mathcal{F}(\overline{X}_t)\|_{\infty} |\overline{V}_t|^2] \ge -\alpha c_{\mathcal{F}} e^{-\alpha \underline{\mathcal{F}}} \mathbb{E}[|\overline{V}_t|^2]. \tag{5.14}$$

We further notice that

$$\begin{aligned}
&\left| \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)} \langle \nabla \mathcal{F}(\overline{X}_t), (X^{\alpha}(\rho_t) - \overline{X}_t) \rangle] \right| \\
&= \left| \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)} \langle \nabla \mathcal{F}(\overline{X}_t) - \nabla \mathcal{F}(X^{\alpha}(\rho_t)), (\overline{X}_t - X^{\alpha}(\rho_t)) \rangle] \right| \\
&\leq e^{-\alpha \underline{\mathcal{F}}} c_{\mathcal{F}} \mathbb{E}[|\overline{X}_t - X^{\alpha}(\rho_t)|^2],
\end{aligned} (5.15)$$

where we have used the fact that $\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}\langle \nabla \mathcal{F}(X^{\alpha}(\rho_t)), (\overline{X}_t - X^{\alpha}(\rho_t))\rangle] = 0$. Furthermore since $\mathbb{E}[|\overline{X}_t - X^{\alpha}(\rho_t)|^2] \leq 2e^{-\alpha \mathcal{F}} \frac{\mathbb{E}[|\delta \overline{X}_t|^2]}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}]}$, one has

$$I_{2} \geq -\frac{\alpha \lambda}{m} e^{-\alpha \underline{\mathcal{F}}} c_{\mathcal{F}} \mathbb{E}[|\overline{X}_{t} - X^{\alpha}(\rho_{t})|^{2}]$$

$$\geq -\frac{\alpha \lambda}{m} 2e^{-2\alpha \underline{\mathcal{F}}} c_{\mathcal{F}} \frac{\mathbb{E}[|\delta \overline{X}_{t}|^{2}]}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}]}.$$
(5.16)

This combining with (5.14) leads to

$$\frac{d^{2}}{dt^{2}}\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{t})}] \geq -\frac{\gamma}{m}\frac{d}{dt}\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{t})}] - \alpha c_{\mathcal{F}}e^{-\alpha\underline{\mathcal{F}}}\mathbb{E}[|\overline{V}_{t}|^{2}]
-\frac{\alpha\lambda}{m}c_{\mathcal{F}}2e^{-2\alpha\underline{\mathcal{F}}}\frac{\mathbb{E}[|\delta\overline{X}_{t}|^{2}]}{\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{t})}]}.$$
(5.17)

Using this, one can obtain

$$\frac{d}{dt}\left(\frac{1}{2}\frac{d}{dt}(\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}])^2\right) = \frac{d}{dt}\left(\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}]\frac{d}{dt}(\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_t)}])\right)$$

$$= \left(\frac{d}{dt} (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}])\right)^{2} + \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}] \frac{d^{2}}{dt^{2}} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}]$$

$$\geq -\frac{\gamma}{2m} \frac{d}{dt} (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}])^{2} - \alpha c_{\mathcal{F}} e^{-2\alpha \underline{\mathcal{F}}} \mathbb{E}[|\overline{V}_{t}|^{2}]$$

$$-\frac{\alpha \lambda}{m} 2e^{-2\alpha \underline{\mathcal{F}}} c_{\mathcal{F}} \mathbb{E}[|\delta \overline{X}_{t}|^{2}]$$

$$\geq -\frac{\gamma}{2m} \frac{d}{dt} (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}])^{2}$$

$$-\left(2\alpha + 2\frac{\alpha \lambda}{m} 2(\frac{2m}{\gamma})^{2}\right) c_{\mathcal{F}} e^{-2\alpha \underline{\mathcal{F}}} \mathbb{E}[\mathcal{H}(t)], \qquad (5.18)$$

where we have used (5.1) in the last inequality. This completes the proof

Our main theorem on global convergence can be described in the following way:

Theorem 5: Under the Assumption 1 and $\mathbf{A1}$, let $(\overline{X}_t, \overline{V}_t)_{t\geq 0}$ be the solution to the nonlinear SDE (4.1). Further we assume that the initial data \overline{X}_0 and \overline{V}_0 satisfy

$$\mu := \frac{\lambda \gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2}\right) \frac{4e^{-\alpha \mathcal{F}}}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}]} > 0, \tag{5.19}$$

and

$$2\alpha \frac{\gamma}{m} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}] \left(\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)} \nabla \mathcal{F}(\overline{X}_0) \cdot \overline{V}_0] \right)_{+} \\ + 4(\alpha + \frac{\alpha \lambda}{m} 2(\frac{2m}{\gamma})^2) c_{\mathcal{F}} e^{-2\alpha \underline{\mathcal{F}}} \frac{\mathbb{E}[\mathcal{H}(0)]}{\chi(\frac{\gamma}{m} - \chi)} < \frac{3}{4} (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}])^2, \quad (5.20)$$

where we denote $x_{+} = \max\{x, 0\}, \forall x \in \mathbb{R}, \text{ and }$

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

Then $\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] \to 0$, $\mathbb{E}[|\overline{V}_t|^2] \to 0$ exponentially fast as $t \to \infty$, and there exists some \widetilde{x} depending on α such that $\mathbb{E}[\overline{X}_t] \to \widetilde{x}$ and $X^{\alpha}(\rho_t) \to \widetilde{x}$ exponentially fast as $t \to \infty$. Moreover it holds that

$$\mathcal{F}(\widetilde{x}) - \underline{\mathcal{F}} \le \frac{1}{\alpha} \log(2) - \frac{1}{\alpha} \log(\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}]) - \underline{\mathcal{F}} \to 0 \text{ as } \alpha \to \infty.$$
 (5.21)

Remark 5.3: If we additionally assume the inverse continuity of \mathcal{F} holds, namely for any $x \in \mathbb{R}^d$ there exists a minimizer x^* of \mathcal{F} (which may depend

on x) such that it holds

$$|x - x^*| < C_0 |\mathcal{F}(x) - \mathcal{F}|^{\ell},$$

where ℓ, C_0 are some positive constants, then one can conclude that $\widetilde{x} \to x^*$ as $\alpha \to \infty$.

Proof: Define

$$T := \inf \left\{ t \ge 0 : \ \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}] < \frac{1}{2} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}] \right\} \text{ with } \inf \emptyset = \infty. \quad (5.22)$$

Obviously, T > 0. Assume that $T < \infty$, then for $t \in [0, T]$, one can deduce that

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

$$\geq \frac{\lambda \gamma}{2m^2} - \left(\frac{2\lambda^2}{\gamma m} + \frac{\sigma^2}{m^2}\right) \frac{4e^{-\alpha \underline{\mathcal{F}}}}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}]} = \mu > 0.$$
(5.23)

Consequently by (5.10) we have

$$\frac{d}{dt}\mathbb{E}[\mathcal{H}(t)] \leq -\frac{\gamma}{m}\mathbb{E}[|\overline{V}_t|^2] - \mu\mathbb{E}[|\delta\overline{X}_t|^2]
\leq -\min\{\mu, \frac{\gamma}{m}\}(\mathbb{E}[|\delta\overline{X}_t|^2] + \mathbb{E}[|\overline{V}_t|^2])
\leq -\frac{\min\{\mu, \frac{\gamma}{m}\}}{\frac{3}{2}((\frac{\gamma}{m})^2 + 1)}\mathbb{E}[\mathcal{H}(t)],$$
(5.24)

where we have used the estimate (5.1). This implies that

$$\mathbb{E}[\mathcal{H}(t)] \le \mathbb{E}[\mathcal{H}(0)] \exp\left(-\frac{\min\{\mu, \frac{\gamma}{m}\}}{\frac{3}{2}((\frac{\gamma}{2m})^2 + 1)}t\right) = \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t). \quad (5.25)$$

One further notice that

$$\chi \le \frac{\frac{\gamma}{m}}{\frac{3}{2}((\frac{\gamma}{2m})^2 + 1)} < \frac{\gamma}{m} .$$

Set $\mathcal{Y}(t) := (\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}])^2$. Then we have

$$\mathcal{Y}'(0) = -2\alpha \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}] \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)} \nabla \mathcal{F}(\overline{X}_0) \cdot \overline{V}_0]. \tag{5.26}$$

By Gronwall's inequality, it follows from Lemma 5.2 that

$$\frac{d}{dt}\mathcal{Y}(t) \ge \mathcal{Y}'(0)\exp(-\frac{\gamma}{m}t) \\ -4\left(\alpha + \frac{\alpha\lambda}{m}2(\frac{2m}{\gamma})^2\right)c_{\mathcal{F}}e^{-2\alpha\underline{\mathcal{F}}}\exp(-\frac{\gamma}{m}t)\int_0^t \exp(\frac{\gamma}{m}s)\mathbb{E}[\mathcal{H}(s)]ds$$

$$\geq \mathcal{Y}'(0) \exp(-\frac{\gamma}{m}t) \\ -4\left(\alpha + \frac{\alpha\lambda}{m}2(\frac{2m}{\gamma})^2\right) c_{\mathcal{F}}e^{-2\alpha\underline{\mathcal{F}}}\mathbb{E}[\mathcal{H}(0)] \exp(-\frac{\gamma}{m}t) \int_0^t \exp((\frac{\gamma}{m} - \chi)s)ds \\ \geq \mathcal{Y}'(0) \exp(-\frac{\gamma}{m}t) \\ -4\left(\alpha + \frac{\alpha\lambda}{m}2(\frac{2m}{\gamma})^2\right) c_{\mathcal{F}}e^{-2\alpha\underline{\mathcal{F}}}\frac{\mathbb{E}[\mathcal{H}(0)]}{\frac{\gamma}{m} - \chi} \exp(-\chi t) ,$$

which implies that

$$\mathcal{Y}(t) \ge \mathcal{Y}(0) - \frac{m}{\gamma} (-\mathcal{Y}'(0))_{+}$$
$$-4 \left(\alpha + \frac{\alpha \lambda}{m} 2(\frac{2m}{\gamma})^{2}\right) c_{\mathcal{F}} e^{-2\alpha \underline{\mathcal{F}}} \frac{\mathbb{E}[\mathcal{H}(0)]}{\chi(\frac{\gamma}{m} - \chi)}.$$

By assumption (5.20), this means that

$$\begin{split} &(\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{t})}])^{2} \\ &\geq (\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{0})}])^{2} - 2\alpha\frac{\gamma}{m}\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{0})}] \left(\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{0})}\nabla\mathcal{F}(\overline{X}_{0})\cdot\overline{V}_{0}]\right)_{+} \\ &- 4\left(\alpha + \frac{\alpha\lambda}{m}2(\frac{2m}{\gamma})^{2}\right)c_{\mathcal{F}}e^{-2\alpha\underline{\mathcal{F}}}\frac{\mathbb{E}[\mathcal{H}(0)]}{\chi(\frac{\gamma}{m} - \chi)} \\ &\geq \frac{1}{4}(\mathbb{E}[e^{-\alpha\mathcal{F}(\overline{X}_{0})}])^{2} \,. \end{split}$$

This means that there exists $\delta > 0$ such that $\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}] \geq \frac{1}{2}\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}]$ in $[T, T + \delta)$ as well. This then contradicts with the definition of T. Hence $T = \infty$. Consequently it holds that

$$\mathbb{E}[\mathcal{H}(t)] \le \mathbb{E}[\mathcal{H}(0)] \exp(-\chi t) \text{ and } \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}] \ge \frac{1}{2} \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}], \quad (5.27)$$

for all $t \geq 0$. Recalling the fact (5.9) this infers that

$$\mathbb{E}[|\overline{X}_{t} - X^{\alpha}(\rho_{t})|^{2}] \leq 2e^{-\alpha \underline{\mathcal{F}}} \frac{\mathbb{E}[|\delta \overline{X}_{t}|^{2}]}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{t})}]}$$

$$\leq 4e^{-\alpha \underline{\mathcal{F}}} (\frac{2m}{\gamma})^{2} \frac{2\mathbb{E}[\mathcal{H}(0)]}{\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_{0})}]} \exp(-\chi t). \tag{5.28}$$

Additionally, one has

$$\mathbb{E}[|\overline{X}_t - \mathbb{E}[\overline{X}_t]|^2] \le 2(\frac{2m}{\gamma})^2 \mathbb{E}[\mathcal{H}(t)]$$

$$\le C \exp(-\chi t) \mathbb{E}[|\overline{V}_t|^2]$$

$$\le 2\mathbb{E}[\mathcal{H}(t)] \le C \exp(-\chi t). \tag{5.29}$$

Moreover we have

$$\left|\frac{d}{dt}\mathbb{E}[\overline{X}_t]\right| \le \mathbb{E}[|\overline{V}_t|] \le C \exp(-\frac{1}{2}\chi t) \to 0 \text{ as } t \to \infty.$$
 (5.30)

This means that $\mathbb{E}[\overline{X}_t] \to \widetilde{x}$ for some \widetilde{x} depending on α , then it follows from (5.29) that $\overline{X}_t \to \widetilde{x}$ in mean square. Thus we have $X^{\alpha}(\rho_t) \to \widetilde{x}$ according to (5.28). Furthermore, by (5.27) one as $\frac{1}{2}\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}] \leq \mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_t)}] \to e^{-\alpha \mathcal{F}(\widetilde{x})}$. Therefore we conclude that

$$\mathcal{F}(\widetilde{x}) \le \frac{1}{\alpha} \log(2) - \frac{1}{\alpha} \log\left(\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}]\right).$$
 (5.31)

By the Laplace principle (3.3), one has

$$0 \le \mathcal{F}(\widetilde{x}) - \underline{\mathcal{F}} \le \frac{1}{\alpha} \log(2) - \frac{1}{\alpha} \log(\mathbb{E}[e^{-\alpha \mathcal{F}(\overline{X}_0)}]) - \underline{\mathcal{F}} \to 0 \text{ as } \alpha \to \infty.$$
 (5.32)

This completes the proof.

6. Numerical examples

In this section, we illustrate through various numerical examples the previous theoretical analysis, i.e., the mean-field limit and the small inertial limit, and analyze the performance of SD-PSO-based methods against various prototype global optimization functions. We refer to [52,57,61,63] for applications of CBO and related methods to high dimensional problems in machine learning.

The SD-PSO algorithm. First we introduce the time discrete versions of the SD-PSO systems [89]. The particle system (2.7) is solved by the *discrete PSO method without local best*

$$\begin{split} X_i^{n+1} &= X_i^n + \Delta t \ V_i^{n+1}, \\ mV_i^{n+1} &= mV_i^n - \gamma \Delta t \ V_i^{n+1} + \lambda \Delta t \left(\overline{X}_{\alpha}^n - X_i^n \right) \\ &+ \sigma \sqrt{\Delta t} \ D(\overline{X}_{\alpha}^n - X_i^n) \ \theta_i^n, \end{split} \tag{6.1}$$

where $\theta_i \sim \mathcal{N}(0,1)$ and the last equation can be rewritten as

$$V_{i}^{n+1} = \left(\frac{m}{m+\gamma \Delta t}\right) V_{i}^{n} + \frac{\lambda \Delta t}{m+\gamma \Delta t} \left(\overline{X}_{\alpha}^{n} - X_{i}^{n}\right) + \frac{\sigma \sqrt{\Delta t}}{m+\gamma \Delta t} D(\overline{X}_{\alpha}^{n} - X_{i}^{n}) \theta_{i}^{n}.$$

$$(6.2)$$

In the general case, the SD-PSO system (3.33) is solved by the discrete PSO method with local best

$$\begin{split} X_i^{n+1} &= X_i^n + \Delta t \ V_i^{n+1}, \\ V_i^{n+1} &= \left(\frac{m}{m+\gamma \ \Delta t}\right) V_i^n + \frac{\lambda_1 \ \Delta t}{m+\gamma \ \Delta t} \left(Y_i^n - X_i^n\right) \\ &+ \frac{\lambda_2 \ \Delta t}{m+\gamma \ \Delta t} \left(\overline{Y}_{\alpha}^n - X_i^n\right) + \frac{\sigma_1 \ \sqrt{\Delta t}}{m+\gamma \ \Delta t} D(Y_i^n - X_i^n) \ \theta_{1,i}^n \\ &+ \frac{\sigma_2 \ \sqrt{\Delta t}}{m+\gamma \ \Delta t} D(\overline{Y}_{\alpha}^n - X_i^n) \ \theta_{2,i}^n, \\ Y_i^{n+1} &= Y_i^n + \nu \ \Delta t \left(X_i^{n+1} - Y_i^n\right) S^{\beta}(X_i^{n+1}, Y_i^n), \end{split} \tag{6.3}$$

where $\theta_{1,i}$, $\theta_{2,i} \sim \mathcal{N}(0,1)$.

Remark 6.1: Note that, the numerical scheme (6.3) using uniform noise becomes equivalent to the PSO algorithm (2.12) under assumptions (2.8) for $\Delta t = 1$, $\nu = 0.5$, and taking the limit α , $\beta \to \infty$ so that Y_i^n , \overline{Y}_{α}^n match the local and global best definitions in (2.3). In addition, in the limit $m \to 0^+$ scheme (6.3) is consistent with the zero-inertia limit (4.35) and reduces to the discrete CBO method with local best

$$X_{i}^{n+1} = X_{i}^{n} + \lambda_{1} \Delta t \left(Y_{i}^{n} - X_{i}^{n} \right) + \lambda_{2} \Delta t \left(\overline{Y}_{\alpha}^{n} - X_{i}^{n} \right)$$

$$+ \sigma_{1} \sqrt{\Delta t} D(Y_{i}^{n} - X_{i}^{n}) \theta_{1,i}^{n} + \sigma_{2} \sqrt{\Delta t} D(\overline{Y}_{\alpha}^{n} - X_{i}^{n}) \theta_{2,i}^{n},$$

$$Y_{i}^{n+1} = Y_{i}^{n} + \nu \Delta t \left(X_{i}^{n+1} - Y_{i}^{n} \right) S^{\beta}(X_{i}^{n+1}, Y_{i}^{n}).$$

$$(6.4)$$

6.1. Validation of the mean field limit

In the following we validate numerically the mean field limit by considering as prototype functions for global optimization the Ackley function and the Rastrigin function in one dimension. The functions have multiple local minima that can easily trap the particle dynamics (see Figure 3). We refer to [68] for additional examples.

The MF-PSO solver. The corresponding MF-PSO equation without local best (3.4) has been discretized using a dimensional splitting where the transport part is solved through a backward semi-Lagrangian method and the remaining Fokker-Planck term is discretized using an implicit central scheme. The MF-PSO equation with memory (3.36) is solved by a further

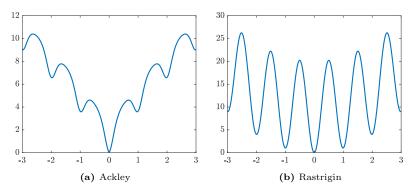


Fig. 3: One-dimensional Ackley and Rastrigin functions in the interval [-3,3] with global minimum in the origin.

dimensional splitting where the additional memory term is discretized using a Lax-Wendroff method. Zero boundary conditions have been implemented outside the computational domain. We refer [68,90] for further details and additional discretizations of Vlasov-Fokker-Planck systems.

In the sequel we used $N=5\times 10^5$ particles, a mesh size for the mean field solver of 90×120 points for $(x,v)\in [-3,3]\times [-4,4]$, and whenever present, the mesh and domain size in y have been taken identical to those in x. To represent the particle solution, we used the probability density estimate based on a normal kernel reconstruction evaluated at equally-spaced points. In all simulations, the initial distribution is assumed to be uniform and the minimum is assumed in x=0.

6.1.1. Absence of memory effects

We consider the optimization process of the Ackley function. Here we report the results obtained with

$$\gamma = 0.5, \quad \lambda = 1, \quad \sigma = 1/\sqrt{3}, \quad \alpha = 30.$$
 (6.5)

The values of λ and σ correspond to the standard PSO choice $c_k = 2$ in (2.8). In Figure 4 we report the contour plots of the evolution, at times t = 0.5, t = 1 and t = 3, of the particle distribution computed through (6.1) and by the direct discretization of the mean-field equation (3.4) together with the evolution in time of the marginal density $\rho(x,t) = \int_{\mathbb{R}^d} f(x,v,t) dv$.

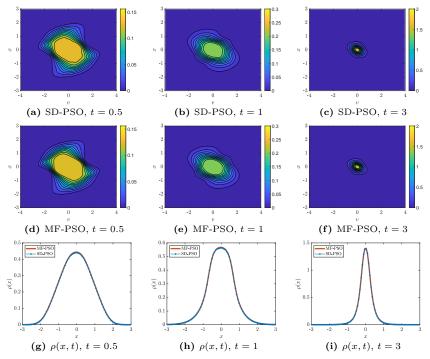


Fig. 4: Mean field validation (no memory). Optimization of the Ackley function. First row: solution of the SD-PSO system (3.1) using $N=5\times 10^5$ particles. Second row: solution of the MF-PSO limit (3.4). Third row: marginal densities.

6.1.2. Only local best dynamics

In the second test case we introduce the dependence from the memory variable and compare the solutions of the discretized stochastic particle model (6.3) with the solver of the mean field limit (3.36) in the case of the Rastrigin function. We assume $\lambda_2 = 0$ and $\sigma_2 = 0$, i.e. only the local best is present. The same parameters (6.5) have been used together with $\beta = 30$ and $\nu = 0.5$ for the local best. In Figure 5 we report the contour plot of the particle and mean-field solutions for the Rastrigin function, where now the final simulation time is t = 6. The corresponding marginal densities are also reported. Also in this second case, one can appreciate the good agreement between the particle and mean-field solutions. We can note that in the presence of local best only, the particles tend to return to their local best position creating a "memory effect" that leads them to concentrate not only in the global minimum but also in the local minima. For large

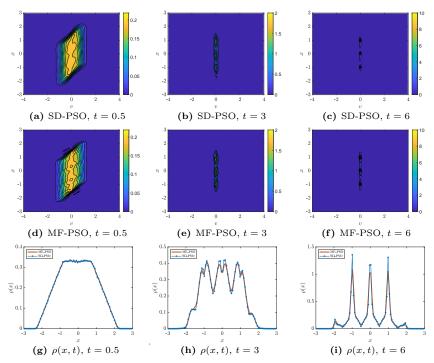


Fig. 5: Mean field validation (local best only). Optimization of the Rastrigin function with minimum in x=0. First row: solution of the SD-PSO system (3.33). Second row: solution of the MF-PSO limit (3.36). Third row: marginal densities.

times we obtain a sequence of particle peaks with zero speed exactly in the positions of the local minima. Thus the dynamic allows us to identify each type of minimum present in the functions.

6.1.3. The general case

In the final test case, we keep the previous scenario, adding the contribution of the global best with the same weight as the local best. Therefore, we take $\lambda_1 = \lambda_2 = 1$, $\sigma_1 = \sigma_2 = 1/\sqrt{3}$ and the same parameters (6.5) in our numerical experiments. The solutions have been obtained by solving the discretized stochastic particle system (6.3) and the deterministic solver of the mean field equation (3.36). In Figure 6 we report the associated marginal density plots. One can observe that the local minima effect disappears and the systems converge consistently towards the global minimum. Note that, by comparing the results for the Ackley function in Figure 6 and

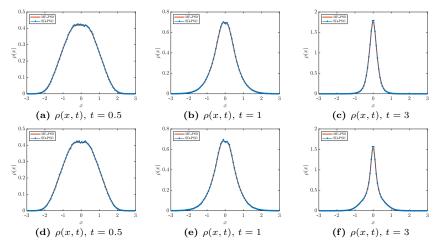


Fig. 6: Mean field validation (general case). Evolution of the density $\rho(x,t)$ of the SD-PSO system (3.33) and the MF-PSO limit (3.36) for two different one-dimensional function with minimum in x=0. First row: optimization on the Ackley function. Second row: optimization on the Rastrigin function.

those in the last row of Figure 4 obtained by solving the same problem in absence of memory terms, at the same time instants, a faster convergence towards the global minimum is observed.

6.2. Numerical small inertia limit

From the analysis in Section 4, the classical CBO model (4.7) is produced as a hydrodynamic approximation of the mean-field PSO system (3.4) in the limit of small inertia. Therefore, we compare the particle solution to a discretization of the mean-field limit CBO system (4.7), starting from the discretization of the stochastic particle model without memory effect (6.1) and decreasing the inertial weight $m \to 0 \ (\gamma \to 1)$.

In Figure 7 we report the plots of the density that describes the solution of the mean-field CBO model and the stochastic PSO model for different inertial weights ($m=0.5,\,m=0.1$ and m=0.01). We considered the minimization problem for the Ackley function with minimum in x=0 and in x=1 with $N=5\times 10^5$ particles for the SD-PSO discretization and a grid of 120 points in space for the mean field CBO solver. It is clear that in the case of m=0.5 the two densities at the final time t=2 are considerably different and a slower convergence is observed in the SD-PSO system, for m=0.1 the agreement is higher and the particle solution seems to converge

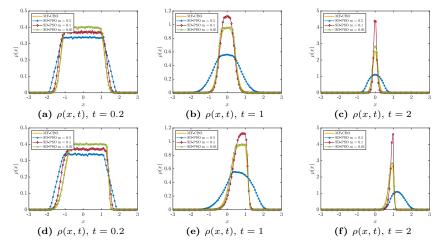


Fig. 7: Low inertia limit. Evolution of the density $\rho(x,t)$ of the SD-PSO discretization (6.1), for decreasing inertial weight m=0.5,0.1,0.01, and the mean-field CBO model (4.7) for the Ackley function with a uniform initial data. First row: minimum in x=0. Second row: minimum in x=1.

faster to the minimum, finally in the case m=0.01 both densities simultaneously grow towards a Dirac delta centered in the minimum. For smaller values of m the two solutions becomes indistinguishable and we omitted the results.

6.3. Performance on high-dimensional test cases

In this section we report the results of several experiments concerning the behavior of the stochastic PSO models, discretized using (6.1) in absence of memory or (6.3) in the general case, in high dimension (d=20) for various prototype test functions (see Table 5). Defining the success rate is critical as it completely alters the performance of the algorithm. In particular, depending on the shape of the objective function, the distance between the estimated minimum and the real minimum can be used as an indicator as in [50–52,60,61]. For some functions, however, this choice may be a poor indicator of the algorithm's performance, since the corresponding value function may be far from its minimum optimal value. In the first round of test cases, since we are limited to the Ackley and Rastrigin functions, for comparison purposes we rely on the choice reported below. Later, when we test the performance of the algorithm for a broader spectrum of test functions, we will generalize the definition of success rate by including the

value of the function [53].

Thus we define:

• the *success rate*, computed averaging over n_r runs and using as convergence criterion

$$\|\overline{X}_{\alpha}^{n_*} - x^*\|_{\infty} < \delta_{err}, \quad \text{or} \quad \|\overline{Y}_{\alpha}^{n_*} - x^*\|_{\infty} < \delta_{err}$$

where x^* is the minimum and n_* the final time.

 \bullet the *error*, evaluated as expected value in the L_2 norm over the successful runs

$$\mathbb{E}(\|\overline{X}_{\alpha}^{n_*} - x^*\|_2), \quad \text{or} \quad \mathbb{E}(\|\overline{Y}_{\alpha}^{n_*} - x^*\|_2);$$

• the number of iterations, where we stop the iteration if

$$\|\overline{X}_{\alpha}^{n} - \overline{X}_{\alpha}^{n-1}\| < \delta_{stall}, \quad \text{or} \quad \|\overline{Y}_{\alpha}^{n} - \overline{X}_{\alpha}^{n-1}\| < \delta_{stall}$$

for n_{stall} consecutive iterations or a maximum n_{max} iterations has been reached.

In the sequel, we consider $n_r = 500$, $\delta_{err} = 0.25$, $\delta_{stall} = 10^{-4}$ and $n_{max} = 10^4$. We remark that, increasing the problem dimension, a larger value of $\alpha \gg 1$ provides better performance [50, 61]. On the other hand, a large value of α may generate numerical instabilities given by the definition of the regularized global best. To avoid this, we used the algorithm presented in [61] which allow the use of arbitrary large values of α .

In the following test cases, we address the role of the various parameters, of the presence of memory and of the local best when solving high dimensional global optimization problems. We refer also to [54] for additional comparisons. In our experiments, the PSO constraints (2.8) have shown strong limitations in terms of success rates and have not been considered. We refer to [68] for further details and comparisons.

6.3.1. Effect of the inertial parameter m

First we test the algorithm performance for the Ackley and the Rastrigin functions in $[-3,3]^d$, d=20. In the left column of Table 1 and Table 2 we report the results obtained without memory effects (6.1) and in the right column the results with memory effects (6.3). Since, typically, optimizing the Rastrigin function is far more difficult than the Ackley function, we explore the space of parameters searching for optimal values of σ and Δt for

Rastrigin			Case w	ithout men	nory		Cas	Case with memory			
\overline{m}		σ	N = 50	N = 100	N = 200	σ_2	N = 50	N = 100	N = 200		
0.00	Rate	9.0	100.0%	100.0%	100.0%	11.0	100.0%	100.0%	100.0%		
	Error		1.19e-04	1.11e-04	9.68 e - 05		6.83e-04	4.70 e-04	4.69e-04		
	n_{iter}		10000.0	10000.0	9912.4		10000.0	9878.2	3290.2		
0.01	Rate	7.0	100.0%	100.0%	100.0%	9.0	100.0%	100.0%	100.0%		
	Error		9.74e-05	2.01e-05	1.62e-05		8.60e-04	8.56e-04	8.81e-04		
	n_{iter}		10000.0	6899.2	2060.1		9939.5	7012.2	5422.1		
0.05	Rate	3.5	37.0%	74.0%	94.0%	4.5	100.0%	100.0%	100.0%		
	Error		4.27e-04	1.26e-04	1.14e-04		1.15e-03	6.67e-04	6.54 e-04		
	n_{iter}		8233.2	7814.0	7326.6		9978.0	7657.6	5639.7		
0.10	Rate	2.0	1.0%	5.5%	29.5%	3.0	80.8%	96.8%	100.0%		
	Error		2.00e-04	1.28e-04	1.11e-04		2.94e-03	8.96 e-04	8.24e-04		
	n_{iter}		6155.4	6221.9	6214.3		9661.5	8676.5	7331.8		

Table 1: SD-PSO with and without memory for $\lambda_1 = \sigma_1 = 0$, $\lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$ and $\alpha = 5 \times 10^4$.

the Rastrigin function, then we used the same values for the Ackley function. This optimization was done empirically through several simulations with simple variations of a given step size for the parameters.

Ackle	y		Case w	ithout men	Case with memory				
\overline{m}		σ	N = 50	N = 100	N = 200	σ_2	N = 50	N = 100	N = 200
0.00	Rate	9.0	100.0%	100.0%	100.0%	11.0	100.0%	100.0%	100.0%
	Error		8.46e-05	4.20 e - 05	1.27 e - 05		1.02e-04	7.66e-05	5.44 e - 05
	n_{iter}		1364.9	1032.4	869.2		2457.0	1778.0	1513.1
0.01	Rate	7.0	100.0%	100.0%	100.0%	9.0	100.0%	100.0%	100.0%
	Error		9.49e-05	5.89 e-05	2.81 e-05		2.34e-03	1.91e-04	1.61e-04
	n_{iter}		2192.9	1886.7	1723.6		6430.4	5447.8	4598.3
0.05	Rate	3.5	100.0%	100.0%	100.0%	4.5	100.0%	100.0%	100.0%
	Error		2.27e-04	1.48e-04	1.03e-04		2.41e-04	1.84e-04	1.48e-04
	n_{iter}		5367.3	4459.4	3928.4		7186.1	5996.0	5074.6
0.10	Rate	2.0	99.5%	100.0%	100.0%	3.0	100.0%	100.0%	100.0%
	Error		8.31e-04	2.76e-04	1.91e-04		3.90e-03	2.64 e-03	2.06e-03
	n_{iter}		5480.8	4514.1	3909.4		8590.6	7326.4	6350.2

Table 2: SD-PSO with and without memory for $\lambda_1 = \sigma_1 = 0$, $\lambda = \lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$ and $\alpha = 5 \times 10^4$.

The results are given for different numbers of particles N. We consider $\alpha = 5 \times 10^4$, whereas the memory parameters β and ν were chosen respectively $\beta = 3 \times 10^3$ and $\nu = 50$. Note that, even if we rely only on the global best since we fix $\lambda_1 = \sigma_1 = 0$, due to the regularization of the memory process the two approaches, with and without memory, differs and a higher

noise in required in presence of memory. Low inertia values yields better performances overall, however, it should be noticed that for the Rastrigin function the best results in term of convergence speed are obtained with a small but non zero inertia value of m=0.01.

6.3.2. Effect of the local best dynamics

Subsequently, we have introduced the local best dynamics in the same optimization process. To reduce the number of free parameters we assume $\lambda_1 = \xi \cdot \lambda_2$, $\sigma = \xi \cdot \sigma_2$ with $\xi \in [0,1]$ so that the local best is always weighted less than the global best. In this test we keep the inertial value m=0 and $\lambda_1=1$, so that we are solving the generalized stochastic differential CBO model with memory using algorithm (6.4). For each value of ξ reported, we have computed an optimal σ_2 achieving the maximum rate of success. We chose $\beta=3\times 10^3$, $\Delta t=0.01$, $\nu=50$ and $\alpha=5\times 10^4$ as in the previous case.

In Tables 3 and 4 we report the behavior of the particle optimizer on the Ackley and Rastrigin functions for different positions of the minimum $x^* = 0$, $x^* = 1$ and $x^* = 2$. Since for large values of ξ we must decrease σ_2 to achieve maximum convergence rate we observe that the total number of iterations may decrease and that a speed-up is obtained thanks to the local best.

Rastrigin		Case $\xi =$	$0, \sigma_2 = 11.$	Case $\xi = 0.25, \sigma_2 = 8.5$			
		N = 50	N = 100	N = 200	N = 50	N = 100	N = 200
$x^* = 0$	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	7.04e-04	4.58e-04	3.29 e-04	9.28e-04	6.11e-04	4.31e-04
	n_{iter}	10000.0	9963.9	4635.1	9978.0	8311.5	5754.1
$x^* = 1$	Rate	98.8%	100.0%	100.0%	99.2%	100.0%	100.0%
	Error	7.08e-04	4.60 e-04	3.27e-04	9.31e-04	6.74 e-04	4.59 e - 04
	n_{iter}	10000.0	10000.0	4670.0	9987.0	9746.7	7460.1
$x^* = 2$	Rate	96.0%	99.1%	100.0%	93.5%	100.0%	100.0%
	Error	6.91 e-04	4.52e-04	3.28e-04	8.78e-04	6.74 e-04	5.66e-04
-	n_{iter}	10000.0	10000.0	5035.5	9980.3	9854.1	8971.9

Table 3: SD-PSO with memory (m=0) for $\lambda_1 = \xi \cdot \lambda_2$, $\sigma = \xi \cdot \sigma_2$, λ_1 , $\lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$, $\alpha = 5 \times 10^4$.

Ackley		Case $\xi =$	$0, \sigma_2 = 11.$	Case $\xi = 0.25, \sigma_2 = 8.5$			
		N = 50	N = 100	N = 200	N = 50	N = 100	N = 200
$x^* = 0$	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	7.36e-05	5.13 e-05	3.26 e- 05	2.54 e-05	1.13e-05	1.07e-05
	n_{iter}	2778.6	2030.0	1623.0	1942.9	1663.8	1442.5
$x^* = 1$	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	7.31e-05	5.14 e - 05	3.26 e- 05	2.58e-05	1.12 e-05	1.02e-05
	n_{iter}	5298.5	3640.6	2575.9	2465.3	1948.5	1632.5
$x^* = 2$	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	7.30e-05	5.07e-05	3.22 e-05	2.64e-05	1.09 e-05	$1.01 e{-}05$
	n_{iter}	7819.8	5771.3	4235.9	3126.8	2286.0	1803.8

Table 4: SD-PSO with memory (m=0) for $\lambda_1 = \xi \cdot \lambda_2$, $\sigma = \xi \cdot \sigma_2$, $\lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$, $\alpha = 5 \times 10^4$.

6.3.3. Comparison on prototype functions

In the last test case we analyze the performance of the methods by solving simultaneously a set of different optimization functions considered in their standard search domains [91] (see Table 5). Here, instead of trying to find an optimal set of parameters for each function we use the same parameters for all functions. Furthermore, in order to identify a comparable set of optimization parameters for the different functions, we found it particularly effective to rescale all functions from their classical domain to the same reference domain. In our experiment we generalized the notion of success criteria by introducing the following definitions

• the *success rate*, computed averaging over n_r runs and using as convergence criterion

$$\|\overline{Y}_{\alpha}^{n_*} - x^*\|_{\infty} < \delta_{err} \quad \text{or} \quad |\mathcal{F}(\overline{Y}_{\alpha}^{n_*}) - \mathcal{F}(x^*)| < \delta_{fun}$$

where x^* is the minimum and n_* the final time.

• The average function value \mathcal{F}_{avg} , computed averaging the function value $\mathcal{F}(\overline{Y}_{\alpha}^{n_*})$ over n_r runs.

In our simulations, we set $[-1,1]^d$ as the reference domain and translate the functions so that all have a minimum value of $\mathcal{F}(x^*)=0$. We selected $\delta_{err}=0.1,\,\delta_{fun}=0.01,\,n_r=500$ and $n_{max}=10^4$. We let most parameters fixed as in previous test case, namely $\alpha=5\times10^4,\,\beta=3\times10^3,\,\nu=50$. Additionally we keep $m=0,\,\Delta t=0.01,$ and for a given value of $\xi=0$ (absence of local best) and $\xi=0.25$ (local best weighted 1/4 of global best) estimate the value for σ_2 in order to maximize the average convergence

Name	Function $\mathcal{F}(x)$	Range	x^*	$\mathcal{F}(x^*)$	Sketch in 2D
Ackley	$-20 \exp \left(-0.2 \sqrt{\frac{1}{d} \sum_{i=1}^{d} (x_i)^2}\right) - \exp \left(\frac{1}{d} \sum_{i=1}^{d} \cos \left(2\pi(x_i)\right)\right) + 20 + e$	$[-32, 32]^d$	(0,,0)	0	
Griewank	$1 + \sum_{i=1}^{d} \frac{(x_i)^2}{4000} - \prod_{i=1}^{d} \cos\left(\frac{x_i}{i}\right)$	$[-600, 600]^d$	(0,,0)	0	
Rastrigin	$10d + \sum_{i=1}^{d} [(x_i)^2 - 10\cos(2\pi(x_i))]$	$[-5.12, 5.12]^d$	(0,,0)	0	2 1 4 4 1 1 5
Rosenbrock	$1 - \cos \left({2\pi \sqrt {\sum\nolimits_{i = 1}^d {(x_i)^2 } } } \right) + 0.1\sqrt {\sum\nolimits_{i = 1}^d {(x_i)^2 } }$	$[-5, 10]^d$	(1,,1)	0	197 6 1 1 1 1 1 1
Salomon	$1 - \cos\left(2\pi\sqrt{\sum_{i=1}^{d}(x_i)^2}\right) + 0.1\sqrt{\sum_{i=1}^{d}(x_i)^2}$	$[-100, 100]^d$	(0,,0)	0	2 4 4 1
Schwefel 2.20	$\sum_{i=1}^{d} x_i $	$[-100, 100]^d$	(0,,0)	0	
XSY random	$\sum_{i=1}^d \eta_i x_i ^i, \qquad \eta_i \sim \mathcal{U}(0,1)$	$[-5, 5]^d$	(0,,0)	0	
XSY 4	$\left(\sum_{i=1}^{d} \sin^2(x_i) - e^{-\sum_{i=1}^{d} (x_i)^2}\right) e^{-\sum_{i=1}^{d} \sin^2 \sqrt{ x_i }}$	$[-10, 10]^d$	(0,,0)	-1	

Table 5: Prototype test functions for global optimization.

rate among all functions. This has been done with simple variations of step 0.5 for σ_2 in the simulations, according to results in Figure 8 where we considered the behavior of the average error and fitness value for different values of σ_2 calculated over n_r runs.

The results in Table 6 confirm the potential of the method in identifying correctly the global minima for different heterogeneous test functions. Overall, with the exception of the Rastrigin function for which the local best produces a reduction in the convergence rate using this set of parameters, the importance of the local best is evident. In particular, the presence of the local best yields a reduction in the number of iterations for the Griewank, the Rosenbrock and the Salomon functions and an increase in

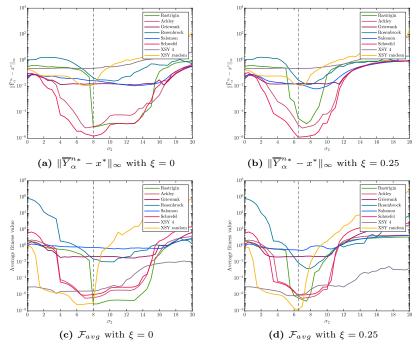


Fig. 8: SD-PSO with memory (m=0). Behavior of the average error (top) and fitness value (bottom) for different values of σ_2 . Here $\sigma_1 = \xi \cdot \sigma_2$, $\lambda_1 = \xi \cdot \lambda_2$, $\lambda_2 = 1$, $\Delta t = 0.01$, $\nu = 50$, $\beta = 3 \times 10^3$ and $\alpha = 5 \times 10^4$. The dashed vertical lines are the estimated optimal values.

the convergence rate for the XSY random and XSY4 functions. Except for the Griewank and Solomon functions, the computed value of the objective function is consistently close to zero and improves by increasing the number of particles. Finally, we emphasize that it was beyond the scope of this survey to discuss additional practical improvements to the algorithms that can be adopted to improve the success rate and the efficiency, like the use of random batch methods [52,79,80], particle reduction techniques [61,62] and parameters adaptivity [4,7]. We refer to [67] for further details on these implementation aspects.

7. Concluding remarks and research directions

PSO methods represent a particularly prominent category within global optimization methods that do not make use of the gradient of the objective function. The popularity of these methods is related to the versatility and

		Case $\xi =$	$0, \sigma_2 = 8.0$)	Case	$\xi = 0.25, \sigma_2$	e = 6.5
		N = 50	N = 100	N = 200	N = 50	N = 100	N = 200
Ackley	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	9.44e-05	3.57e-05	1.48e-05	9.25e-06	4.40e-06	2.02e-06
	\mathcal{F}_{avg}	2.61e-05	1.04 e-05	8.49 e-06	2.65e-05	1.26e-05	5.78e-06
	n_{iter}	1012.5	847.9	736.2	1033.4	874.3	764.0
Griewank	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	2.28e-02	2.24e-02	2.19e-02	2.27e-02	2.16e-02	2.24e-02
	\mathcal{F}_{avg}	5.57e-02	5.21e-02	4.26e-02	5.25 e-02	4.93e-02	2.28e-02
	n_{iter}	1010.8	861.6	761.7	1006.3	734.7	626.6
Rastrigin	Rate	34.0%	70.7%	95.0%	9.0%	26.4%	42.0%
	Error	1.78e-05	1.89e-05	2.05e-05	3.01e-05	3.12e-05	3.03 e-05
	\mathcal{F}_{avg}	9.32e-08	9.68e-08	9.95 e-08	2.41e-07	2.58e-07	2.44e-07
	n_{iter}	1308.5	1122.9	970.5	1631.0	1483.0	1334.8
Rosenbrock	Rate	49.3%	84.7%	100.0%	87.3%	100.0%	100.0%
	Error	2.60e-02	3.44e-02	1.08e-02	4.87e-02	3.32e-02	6.92 e-03
	\mathcal{F}_{avg}	8.58 e-02	1.25 e-02	9.30e-03	2.12e-02	8.01 e-03	3.23 e-04
	n_{iter}	8009.3	8392.8	7358.0	9669.8	9553.8	7925.7
Schwefel 2.20	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	2.11e-05	1.73 e-06	7.32e-07	3.65e-06	1.63e-06	1.09e-06
	\mathcal{F}_{avg}	2.93 e-03	4.99e-04	2.18e-04	5.14e-05	2.46 e - 05	8.01 e-06
	n_{iter}	865.9	749.8	668.3	863.2	747.0	665.8
Salomon	Rate	84.7%	98.7%	100.0%	100.0%	100.0%	100.0%
	Error	8.94 e-02	6.45 e - 02	4.99e-02	3.72e-02	3.21 e-02	2.75 e-02
	\mathcal{F}_{avg}	8.96 e-01	6.66e-01	5.24 e-01	3.83e-01	3.21 e-01	2.75 e - 01
	n_{iter}	1749.3	1657.9	1631.9	2193.7	1749.7	1138.2
XSY random	Rate	90.0%	99.3%	100.0%	100.0%	100.0%	100.0%
	Error	4.11e-02	2.26e-02	1.14 e-02	2.45 e-02	1.67e-02	1.66e-02
	\mathcal{F}_{avg}	5.64 e-07	9.60 e-08	6.06e-08	9.75e-09	7.26e-09	4.56e-09
	n_{iter}	10000.0	10000.0	10000.0	10000.0	10000.0	10000.0
XSY 4	Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Error	1.09e+00	9.85 e - 01	9.70 e-01	8.56e-01	8.19 e-01	7.97e-01
	\mathcal{F}_{avg}	2.88 e - 05	2.57e-05	7.44e-05	1.69e-07	1.42 e-07	1.41 e - 07
	n_{iter}	9682.5	9018.1	8861.6	10000.0	10000.0	10000.0

Table 6: SD-PSO with memory (m=0) for $\lambda_1=\xi\cdot\lambda_2,\ \sigma_1=\xi\cdot\sigma_2,\ \lambda_2=1,\ \Delta t=0.01,\ \nu=50,\ \beta=3\times10^3,\ \alpha=5\times10^4.$

robustness of the algorithms, the good scalability that allows dealing with high-dimensional problems, and the ability to identify the global minimum effectively even in the case of non-convex and possibly non-smooth functions. Despite this, a complete mathematical theory related to the derivation of such methods and their global convergence properties is still lacking.

In this work, relying on some recent results [67, 68, 70–72], we have made an important step towards the construction of a general mathematical theory for the rigorous analysis and the understanding of PSO methods. The starting point of our analysis is a generalization of PSO methods in the context of second-order stochastic differential equations. In addition

to the continuous formulation of PSO algorithms this novel class of methods generalizes the particle optimization process by making the alignment and exploration coefficients, based on the corresponding drift and diffusion dynamics, independent.

In the mean-field limit, using a regularized version of these SD-PSO systems, we obtained a Vlasov-Fokker-Planck type equation describing the MF-PSO dynamics. In addition, we rigorously studied the behavior of the system for small values of the inertia parameter showing how in such a limit the MF-PSO dynamics converges to a generalization of CBO models containing the local best. The latter result allowed us to clarify the relationships between these two classes of meta-heuristic optimization methods. A convergence result to the global minimum for a wide class of objective function is then proved in the case where the dynamic does not take into account memory effects. A complete gallery of numerical examples illustrate on the one hand the theoretical results obtained and on the other hand how the new class of SD-PSO methods potentially presents several advantages over traditional PSO in terms of convergence speed and solution stability.

These results open important perspectives in the area of mathematical understanding of particle swarming optimization methods and in the construction of new algorithms. Among the many research directions some, not exhaustive, are summarized below.

- The majority of PSO applications are limited to single objective and unconstrained optimization problems. Therefore, the development of methods capable to deal with multi-objective and constrained optimization problems is a challenging and interesting area of research.
- Most of the convergence results for mean-field PSO and CBO models refer to the global best only. Generalization of these results to include the effect of the local best and its role should be studied. Convergence rate estimates of practical interest are still limited and further analysis is necessary.
- Similarly to classical PSO algorithms, the computational parameters are usually determined according to specific problems and require considerable application experience and numerous experimental tests. The identification of optimal parameters and the implementation of adaptive techniques for their determination is thus fundamental for many applications.

Acknowledgments

This work has been written within the activities of GNCS group of INdAM (National Institute of High Mathematics). The support of MIUR-PRIN Project 2017, No. 2017KKJP4X "Innovative numerical methods for evolutionary partial differential equations and applications" and of the ESF PhD grant "Mathematical and statistical methods for machine learning in biomedical and socio-sanitary applications" is acknowledged. H. H. is partially supported by the Pacific Institute for the Mathematical Sciences (PIMS) postdoc fellowship. J. Q. is partially supported by the National Science and Engineering Research Council of Canada (NSERC) and by the start-up funds from the University of Calgary.

References

- J. Kennedy and R. Eberhart. Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks, vol. 4, pp. 1942–1948, 1995.
- J. Kennedy. The particle swarm: social adaptation of knowledge. In Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97), pp. 303-308, 1997.
- J. Kennedy, Particle Swarm Optimization, In Encyclopedia of Machine Learning. Springer US, Boston, MA, 2010.
- 4. R. Poli, J. Kennedy, and T. Blackwell, Particle swarm optimization, *Swarm intelligence*. **1** (2007), 33–57.
- 5. Y. Shi and R. Eberhart. A modified particle swarm optimizer. In *International conference on evolutionary computation proceedings. IEEE world congress on computational intelligence*, pp. 69–73, 1998.
- H. M. Emara and H. A. A. Fattah. Continuous swarm optimization technique with stability analysis. In *Proceedings of the 2004 American control conference*, vol. 3, pp. 2811–2817, 2004.
- 7. D. Wang, D. Tan1, and L. Liu, Particle swarm optimization algorithm: an overview, *Soft Comput.* (2017).
- 8. I. Aoki, A simulation study on the schooling mechanism in fish, Bull. Jpn. Soc. Sci. Fish. 48 (1982), 1081–1088.
- 9. A. Okubo, Dynamical aspects of animal grouping: swarms, schools, flocks, and herds, *Adv. Biophys.* **22** (1986), 1–94.
- T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Novel type of phase transition in a system of self-driven particles, *Physical Review Letters*. 75 (1995), 1226–1229.
- 11. I. Giardina, Collective behavior in animal groups: Theoretical models and empirical studies, *HFSP Journal.* **2** (2008), 205–219.
- 12. C. Reynolds. Flocks, herds and schools: A distributed behavioral model. In Proceedings of the 14th Annual Conference on Computer Graphics and Inter-

- active Techniques (SIGGRAPH '87). Association for Computing Machinery, pp. 25–34, 1987.
- 13. J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, pp. 297–336. Springer, 2010.
- 14. D. Sumpter, Collective animal behavior. Princeton University Press, 2010.
- E. Bonabeau, M. Dorigo, and G. Theraulaz, Swarm Intelligence: From Natural to Artificial System. Oxford University Press, 1999.
- F. Cucker and S. Smale, Emergent behavior in flocks, IEEE Trans. Automat. Control. 52 (2007), 852–862.
- S. Motsch and E. Tadmor, Heterophilious dynamics enhances consensus, SIAM Rev. 56 (2014), 577–621.
- 18. Y.-P. Choi and S. Salem, Cucker-Smale flocking particles with multiplicative noises: Stochastic mean-field limit and phase transition, *Kinetic & Related Models.* **12** (2019), 573–592.
- T. Bäck, D. B. Fogel, and Z. Michalewicz, Handbook of Evolutionary Computation. IOP Publishing Ltd., 1997.
- C. Blum and A. Roli, Metaheuristics in combinatorial optimization: Overview and conceptual comparison, ACM computing surveys (CSUR). 35 (2003), 268–308.
- M. Gendreau, J. Y. Potvin, et al., Handbook of metaheuristics. vol. 2, Springer, 2010.
- 22. W. Wong and C. I. Ming. A review on metaheuristic algorithms: Recent trends, benchmarking and applications. In 2019 7th International Conference on Smart Computing Communications (ICSCC), pp. 1–5, 2019. doi: 10.1109/ICSCC.2019.8843624.
- J. Larson, M. Menickelly, and S. M. Wild, Derivative-free optimization methods, Acta Numerica. 28 (2019), 287–404.
- A. Neumaier, Complete search in continuous global optimization and constraint satisfaction, Acta Numerica. 13 (2004), 271–369.
- C. Audet and W. L. Hare, Derivative-Free and Blackbox Optimization. Springer, 2017.
- 26. R. Holley and D. Stroock, Simulated annealing via Sobolev inequalities, Communications in Mathematical Physics. 115 (1988), 553–569.
- S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Optimization by simulated annealing, Science. 220 (1983), 671–680.
- 28. E. Aarts and J. Korst, Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing. John Wiley & Sons, Inc., 1989.
- 29. M. Dorigo, V. Maniezzo, and A. Colorni, Ant system: Optimization by a colony of cooperating agents, *IEEE Transactions on Systems, Man, and Cybernetics Part B.* **26** (1996), 29–41.
- 30. M. Dorigo and C. Blum, Ant colony optimization theory: A survey, *Theoretical computer science*. **344** (2005), 243–278.
- 31. J. H. Holland et al., Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.

- MIT press, 1992.
- 32. D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley Professional, 1989.
- D. B. Fogel, Evolutionary computation: toward a new philosophy of machine intelligence. vol. 1, John Wiley & Sons, 2006.
- R. Storn and K. Price, Differential evolution a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimiza*tion. 11 (1997), 341–359.
- 35. K. Sörensen, Metaheuristics—the metaphor exposed, International Transactions in Operational Research. 22 (2015), 3–18.
- V. Bruned, A. Mas, and S. Włodarczyk, Weak convergence of particle swarm optimization, arXiv preprint arXiv:1811.04924. (2019).
- B. I. Schmitt, Convergence Analysis for Particle Swarm Optimization. FAU University Press, 2015.
- G. Xu and G. Yu, On convergence analysis of particle swarm optimization algorithm, *Journal of Computational and Applied Mathematics*. 340 (2018), 709–717.
- N. Bellomo and S. Y. Ha, A quest toward a mathematical theory of the dynamics of swarms, Mathematical Models and Methods in Applied Sciences. 27 (2017), 745–770.
- R. Poli, Mean and variance of the sampling distribution of particle swarm optimizers during stagnation, *IEEE Transactions on Evolutionary Computa*tion. 13 (2009), 712–721.
- Y. Zhang, S. Wang, and G. Ji, A comprehensive survey on particle swarm optimization algorithm and its applications, *Mathematical problems in engi*neering. 2015 (2015).
- P.-E. Jabin and Z. Wang. Mean field limit for stochastic particle systems. In Active Particles, Volume 1, pp. 379–402. Springer, 2017.
- A. S. Sznitman. Topics in propagation of chaos. In Ecole d'été de probabilités de Saint-Flour XIX—1989, pp. 165–251. Springer, 1991.
- 44. F. Golse, The mean-field limit for the dynamics of large particle systems, Journées équations aux dérivées partielles. (2003), 1–47.
- P.-E. Jabin, A review of the mean field limits for Vlasov equations, Kinetic & Related Models. 7 (2014), 661.
- J. A. Carrillo and Y.-P. Choi, Mean-field limits: From particle descriptions to macroscopic equations, Arch. Ration. Mech. Anal. 241 (2021), 1529–1573.
- J. A. Carrillo, Y.-P. Choi, and S. Salem, Propagation of chaos for the Vlasov-Poisson-Fokker-Planck equation with a polynomial cut-off, *Commun. Contemp. Math.* 21 (2019), 1850039, 28.
- H. Huang, J.-G. Liu, and P. Pickl, On the mean-field limit for the Vlasov-Poisson-Fokker-Planck system, J. Stat. Phys. 181 (2020), 1915–1965.
- F. Bolley, J. A. Canizo, and J. A. Carrillo, Stochastic mean-field limit: non-Lipschitz forces and swarming, Mathematical Models and Methods in Applied Sciences. 21 (2011), 2179–2210.
- R. Pinnau, C. Totzeck, O. Tse, and S. Martin, A consensus-based model for global optimization and its mean-field limit, Mathematical Models and

- Methods in Applied Sciences. 27 (2017), 183–204.
- 51. J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse, An analytical framework for consensus-based global optimization method, *Mathematical Models and Methods in Applied Sciences.* **28** (2018), 1037–1066.
- J. A. Carrillo, S. Jin, L. Li, and Y. Zhu. A consensus-based global optimization method for high dimensional machine learning problems. In ESAIM: Control, Optimisation and Calculus of Variations, 2020.
- C. Totzeck and M. T. Wolfram, Consensus-based global optimization with personal best, Mathematical Biosciences and Engineering. 17 (2020), 6026– 6044.
- 54. C. Totzeck, R. Pinnau, S. Blauth, and S. Schotthöfer, A numerical comparison of consensus-based global optimization to other particle-based global optimization schemes, *PAMM*. **18** (2018), 1–28.
- S. Y. Ha, S. Jin, and D. Kim, Convergence of a first-order consensus-based global optimization algorithm, Mathematical Models and Methods in Applied Sciences. 30 (2020), 2417–2444.
- S. Y. Ha, S. Jin, and D. Kim, Convergence and error estimates for timediscrete consensus-based optimization algorithms, *Numerische Mathematik*. 147 (2021), 255–282.
- 57. J. Chen, S. Jin, and L. Lyu, A consensus-based global optimization method with adaptive momentum estimation, arXiv preprint arXiv:2012.04827.
- 58. M. Fornasier, T. Klock, and K. Riedl, Consensus-based optimization methods converge globally in mean-field law, arXiv preprint arXiv:2103.15130. (2021).
- 59. C. Totzeck, Trends in consensus-based optimization, arXiv preprint arXiv:2104.01383. (2021).
- 60. M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen, Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit, *Mathematical Models and Methods in Applied Sciences.* **30** (2020), 2725–2751.
- 61. M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen, Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning, *J. Machine Learning Research*. (2021), to appear.
- 62. M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen, Anisotropic diffusion in consensus-based optimization on the sphere, arXiv preprint arXiv:2104.00420. (2021).
- 63. A. Benfenati, G. Borghi, and L. Pareschi, Binary interaction methods for high dimensional global optimization and machine learning, arXiv preprint arXiv:2105.02695. (2021).
- 64. L. Pareschi and G. Toscani, Interacting multiagent systems: kinetic equations and Monte Carlo methods. OUP Oxford, 2013.
- 65. G. Naldi, L. Pareschi, and G. Toscani, eds., Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences, vol. 10, Modeling and Simulation in Science, Engineering and Technology. Birkhäuser Basel, 2010.
- J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes, Consensus based sampling, arXiv preprint arXiv:2106.02519. (2021).
- 67. S. Grassi and L. Pareschi, Mean field particle swarm optimization and consen-

- sus based optimization: fast algorithms and implementation aspects, preprint. (2021).
- 68. S. Grassi and L. Pareschi, From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit, Mathematical Models and Methods in Applied Sciences. 0 (2021), 1–33.
- H. Huang, A note on the mean-field limit for the particle swarm optimization, Applied Mathematics Letters. 117 (2021), 107133.
- H. Huang and J. Qiu, On the mean-field limit for the consensus-based optimization, arXiv preprint arXiv:2105.12919. (2021).
- C. Cipriani, H. Huang, and J. Qiu, Zero-inertia limit: from particle swarm optimization to consensus based optimization, arXiv preprint arXiv:2104.06939. (2021).
- H. Huang, J. Qiu, and K. Riedl, On the global convergence of particle swarm optimization methods, preprint. (2021).
- 73. J.-A. Acebrón and R. Spigler, Adaptive frequency model for phase-frequency synchronization in large populations of globally coupled nonlinear oscillators, *Physical Review Letters.* **81** (1998), 2229–2232.
- M. H. Duong, A. Lamacz, M. A. Peletier, and U. Sharma, Variational approach to coarse-graining of generalized gradient flows, Calculus of variations and partial differential equations. 56 (2017), 65–100.
- J. H. M. Evers, R. C. Fetecau, and W. Sun, Small inertia regularization of an anisotropic aggregation model, *Math. Models Methods Appl. Sci.* 27 (2017), 1795–1842.
- R. C. Fetecau and W. Sun, First-order aggregation models and zero inertia limits, J. Diff. Equations. 259 (2015), 6774–6802.
- 77. Y.-P. Choi and O. Tse, Quantified overdamped limit for kinetic Vlasov-Fokker-Planck equations with singular interaction forces, arXiv preprint arXiv:2012.00422. (2020).
- 78. J. A. Carrillo, Y.-P. Choi, and Y. Peng, Large friction-high force fields limit for the nonlinear Vlasov–Poisson–Fokker–Planck system, arXiv preprint arXiv:2103.12276. (2021).
- 79. G. Albi and L. Pareschi, Binary interaction algorithms for the simulation of flocking and swarming dynamics, *Multiscale Modeling & Simulation*. 11 (2013), 1–29.
- S. Jin, L. Li, and J. G. Liu, Random Batch Methods (RBM) for interacting particle systems, *Journal of Computational Physics*. 400 (2020), 108877.
- 81. C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.
- V. Vapnik. Principles of risk minimization for learning theory. In Advances in neural information processing systems, pp. 831–838, 1991.
- P. Jain and P. Kar, Non-convex optimization for machine learning, Foundations and Trends in Machine Learning. 10 (2017), 142–363.
- 84. R. Durrett, Stochastic calculus: a practical introduction. CRC press, 2018.
- 85. P. Billingsley, Convergence of probability measures. John Wiley & Sons, 2013.
- H. P. McKean Jr, A class of Markov processes associated with nonlinear parabolic equations, Proceedings of the National Academy of Sciences of the United States of America. 56 (1966), 1907.

- 87. G. Da Prato and J. Zabczyk, Stochastic equations in infinite dimensions. Cambridge university press, 2014.
- 88. X. Mao, Stochastic differential equations and applications. Elsevier, 2007.
- 89. E. Platen, An introduction to numerical methods for stochastic differential equations, *Acta numerica*. **8** (1999), 197–246.
- G. Dimarco and L. Pareschi, Numerical methods for kinetic equations, Acta Numerica. 23 (2014), 369–520.
- 91. M. Jamil and X. S. Yang, A literature survey of benchmark functions for global optimisation problems, *International Journal of Mathematical Modelling and Numerical Optimisation*. 4 (2013), 150–194.