SWARM-BASED GRADIENT DESCENT MEETS SIMULATED ANNEALING*

ZHIYAN DING†, MARTIN GUERRA‡, QIN LI‡, AND EITAN TADMOR§

Abstract. We introduce a novel method, called swarm-based simulated annealing (SSA), for nonconvex optimization which is at the interface between the swarm-based gradient-descent (SBGD) [J. Lu et al., arXiv:2211.17157; E. Tadmor and A. Zenginoglu, *Acta Appl. Math.*, 190 (2024)] and simulated annealing (SA) [V. Cerny, *J. Optim. Theory Appl.*, 45 (1985), pp. 41–51; S. Kirkpatrick et al., *Science*, 220 (1983), pp. 671–680; S. Geman and C.-R. Hwang, *SIAM J. Control Optim.*, 24 (1986), pp. 1031–1043]. Similarly to SBGD, we introduce a swarm of agents, each identified with a position, $\mathbf{x}$ *and* mass m, to explore the ambient space. Similarly to SA, the agents proceed in the gradient descent direction, and are subject to Brownian motion. The annealing rate, however, is dictated by a decreasing function of their mass. As a consequence, instead of the SA protocol for time-decreasing temperature, here the swarm decides how to "cool down" agents, depending on their own accumulated mass. The dynamics of masses is coupled with the dynamics of positions: agents at higher ground transfer (part of) their mass to those at lower ground. Consequently, the resulting SSA optimizer is dynamically divided between heavier, cooler agents viewed as "leaders" and lighter, warmer agents viewed as "explorers." Mean-field convergence analysis and benchmark optimizations demonstrate the effectiveness of the SSA method as a multidimensional global optimizer.

Key words. optimization, stochastic gradient descent, swarm-based gradient descent, provisional minimum, simulated annealing

MSC codes. 90C26, 65K10, 92D25

DOI. 10.1137/24M1657808

1. Introduction. We introduce a new swarm-based optimization method to compute the global minimum of nonconvex objective functions, $\mathbf{x}_* := \mathrm{argmin}_{\Omega \subset \mathbb{R}^d} F(\mathbf{x})$. The swarm consists of N agents—enumerated $j = 1, 2, \ldots, N$, each of which is identified with a time-dependent position, $\mathbf{x}_t^j \in \mathbb{R}^d$, and time-dependent mass, $m_t^j \in \mathbb{R}_+$. The positions are governed by an overdamped Langevin process,[1]

$$\mathrm{d}\mathbf{x}_t^j = -\nabla F\left(\mathbf{x}_t^j\right)\mathrm{d}t + \sqrt{2\sigma_t^j}\mathrm{d}W_t^j, \qquad j = 1, 2, \ldots, N,$$

where $\{W_t^j\}_{j=1}^N$ are independent Brownian motions with amplitude $\sqrt{2\sigma_t^j}$, where σ_t^j is the scaled "temperature" (dating back to Langevin [20]), or are referred to as the

---

†Department of Mathematics, University of California, Berkeley, CA 94720 USA (zding.m@berkeley.edu).
‡Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 USA (mguerra4@wisc.edu, qinli@math.wisc.edu).
§Department of Mathematics and Institute for Physical Science & Technology University of Maryland, College Park, MD 20742 USA (tadmor@umd.edu).
[1]A word about notation. Throughout the paper the time-dependent quantities in deterministic models are denoted $\square(t,\ldots)$ while time-dependent quantities which arise from stochastic descriptions are denoted $\square_t(\ldots)$.

annealing rate, [18, 5, 12, 11, 7]. The intricate aspect of the dynamics is a proper tuning of these annealing-rates σ_t^j. This is where the masses, $\{m_t^j\}$, come into play: they are driven by

$$\mathrm{d}m_t^j = -m_t^j \left(F\left(\mathbf{x}_t^j\right) - \overline{F}_t^N \right) \mathrm{d}t, \qquad j = 1, 2, \ldots, N,$$

where $\overline{F}_t^N$ is the mass-weighted average of the N agents, $\overline{F}_t^N := \frac{\sum_{j=1}^N m_t^j F(\mathbf{x}_t^j)}{\sum_{j=1}^N m_t^j}$. We tune the annealing rates as a decreasing function of the mass, $\sigma_t^j = \sigma(m_t^j)$. Thus, our protocol for annealing rate prevents lighter agents from being trapped in basins of *local* minima while it "cools down" heavier agents at the basin of attraction of the *global* minimum of F. This is where the swarm-based gradient descent [21] meets simulated annealing (SA), [18, 10, 5], except that the protocol for the "cooling process" is different: instead of an explicit recipe for decreasing the temperature as is done in SA, our swarm-based approach lets different agents at a lower ground cool down, depending on their increasing mass $\sigma_t^j = \sigma(m_t^j)$.

Why provisional minimum? We refer to the mass-weighted average

$$(1.1) \qquad \overline{F}_t^N := \frac{\sum_{j=1}^N m_t^j F(\mathbf{x}_t^j)}{\sum_{j=1}^N m_t^j}$$

as the *provisional minimum.* To clarify why a provisional minimum is needed, we note that according to the mass equation above, agents that are above the provisional minimum will shed a fraction of their mass—masses that are transferred to agents that are below the provisional minimum. Since the total mass remains constant, $\sum_j \mathrm{d}m_t^j/\mathrm{d}t = 0$, masses tend to concentrate with agents near or below the provisional minimum, and it therefore makes sense to adjust the annealing rate, σ_t^j, as a *decreasing* function of the mass,

$$\sigma_t^j = \sigma(m_t^j) : (0, \infty) \mapsto \mathbb{R}_+, \qquad j = 1, 2, \ldots, N.$$

Given this adjustment of mass-dependent annealing, *it is expected* that for a large crowd of agents, $N \gg 1$, (1.1) will approach the global minimum,

$$\lim_{N \to \infty} \overline{F}_t^N \overset{t \to \infty}{\longrightarrow} F_* := F(\mathbf{x}_*).$$

Indeed, our main result states that, under the appropriate assumptions (in particular, the assumption of uniqueness of the global minimizer made in Assumption 2.3 below), this is true.

The method described above lies at the interface of the deterministic swarm-based gradient descent (SBGD) approach and the stochastic-based SA. We recall that the SBGD is governed by a deterministic system which governs the swarm of N-agents, and for $j = 1, 2, \ldots, N$,

$$(\text{SBGD}) \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^j(t) = -\alpha(m^j(t))\nabla F(\mathbf{x}^j(t)), \quad \frac{\mathrm{d}}{\mathrm{d}t}m^j(t) = -\left(F(\mathbf{x}^j(t)) - \overline{F}^N(t)\right)m^j(t).$$

The key feature advocated in (SBGD) is swarm dynamics of both positions and masses embedded in $\mathbb{R}^d \times \mathbb{R}_+$. The additional dimension of mass serves as a platform for *communication among agents,* encoded in the provisional minimum, which in [21] was taken as the minimum of the crowd at the given time, $\overline{F}^N(t) = \min_j F(\mathbf{x}^j(t))$.

The same swarm-based methodology is used in this work, except for two distinct features. First, the provisional minimum is given here by (1.1) which eventually is expected to approach the global minimum, $\min_j F(\mathbf{x}^j(t))$, used in [21]. The second and more essential distinction, is how masses are being used to adjust the dynamics of positions along the gradient descent: in SBGD, it is the *time step* which is adjusted as a decreasing function of the mass, $\alpha = \alpha(m^j(t))$. In our new proposed algorithm, masses are used to adjust the annealing rate, $\sigma = \sigma(m_t^j)$, as a decreasing function of mass. This new algorithm is named swarm-based simulated annealing (SSA) optimizer:

$$
\text{(SSA)} \qquad
\begin{cases}
\mathrm{d}\mathbf{x}_t^j = -\nabla F(\mathbf{x}_t^j)\mathrm{d}t + \sqrt{2\sigma(m_t^j)}\mathrm{d}W_t^j, \\
\mathrm{d}m_t^j = -\big(F(\mathbf{x}_t^j) - \overline{F}_t^N\big)m_t^j\mathrm{d}t,
\end{cases}
\qquad j = 1, 2, \ldots, N.
$$

Remark 1.1 (on the choice of provisional minimum). The SBGD method (SBGD) employs the actual minimum, $\overline{F}(t) = \min_j F(\mathbf{x}^j(t))$ as its provisional minimum [21, 29]. This, however, is not amenable to the mean-field limit analysis that we will pursue in section 2.1. The mean-field limit requires us to tour the whole landscape, ending with

$$
\lim_{N \to \infty} \overline{F}^N(t) = \lim_{N \to \infty} \min_{j=1,\ldots,N} F(\mathbf{x}^j(t)),
$$

which is already the global minimum we are looking for. Instead, the use of mass-weighted average $\overline{F}_t^N$ in (SSA) as a provisional minimum, admits a mean-field interpretation which is shown to be driven towards the same desired global minimum.

1.1. In-swarm communication combined with stochastic search. Application of deterministic swarm-based methods reveals that such methods *succeed in the accuracy sense but fail in the probability sense,*—namely, when these methods succeed, then they find the global optimizer, yet there is a nonzero probability for the methods to fail to do so. In contrast, stochastic methods *succeed in the probability sense but fail in the accuracy sense,*—namely, such methods can always find the global basin, but they do not necessarily return a true global optimizer. In this context, our swarm-based optimization with adjusted annealing rate takes advantage of randomness in $(\text{SSA})_1$ combined with communication-based swarming in $(\text{SSA})_2$, with the aim of succeeding in both the probability and the accuracy sense. This is achieved by a dynamic process that combines swarm communication and randomness, as described below.

The essential role of communication in SBGD was already emphasized in [21, 29]. This is amplified in the present context, upon setting $\sigma(\cdot) \equiv 0$ in $(\text{SSA})_1$. One ends up with a crowd of N independent agents driven by gradient descent,

$$
\dot{\mathbf{x}}^j(t) = -\nabla F(\mathbf{x}^j(t)), \qquad j = 1, 2, \ldots, N.
$$

The success of such a noncommunicating crowd in exploring the ambient landscape is *significantly worse* than that of a communication-based crowd. In section 4, we show that the method fails in the probability sense.

Communication is encoded in the provisional minimum, where agents "communicate" their height relative to $\overline{F}_t^N$. In our swarm-based optimization with adjusted annealing rate (SSA), $\overline{F}_t^N$ is taken as the weighted average height, weighted by the different masses. Consequently, there is a dynamic distinction between agents above the provisional minimum and agents below or at the level of the provisional minimum.

Agents above the provisional minimum carry lighter mass and therefore explore the ambient landscape with a relatively large Brownian motion; these are viewed as the *explorers* of the crowd. The heavier agents—those that are below or at the level of the provisional minimum, evolve in the gradient direction with a smaller Brownian motion and are viewed as *leaders* of the crowd. Of course, once an exploring light agent "hits" a new lower ground at, say, $\mathbf{x}_t^j$, where $F(\mathbf{x}_t^j) \ll -\overline{F}_t^N$, it is expected to accumulate more mass from the higher agents, to cool down, and eventually to become a leader, driving the crowd to a new lower point.

We note that by fixing a constant annealing rate, $(\text{SSA})_1$ with $\sigma(m_t^j) \equiv \sigma > 0$, we also encounter a problem. In this case, we recover the Langevin Monte Carlo

$$\mathrm{d}\mathbf{x}_t^j = -\nabla F(\mathbf{x}_t^j)\mathrm{d}t + \sqrt{2\sigma}\mathrm{d}W_t^j, \qquad j = 1, 2, \ldots, N,$$

whose long time invariant measure suggests $\mathbf{x} \propto e^{-F(\mathbf{x})/\sigma}$. It admits the proper convergence in the zero-variance limit, $\mathbf{x} \xrightarrow{\sigma \to 0} \mathrm{argmin}_{\mathbf{x}} F(\mathbf{x})$, but, as we show in section 4 below, the resulting method fails in the accuracy sense for any finite σ. Thus, the dynamic process of relabeling explorers and leaders, coupled through a mass-dependent annealing rate, $\sigma_t^j = \sigma(m_t^j)$, is at the heart of the matter.

In section 5 we prove global convergence of (SSA) in both the sense of probability and the sense of accuracy. This is achieved in a two-step argument: first, we conduct a large-crowd, mean-field analysis summarized in Theorem 2.2 below, which allows us to pass from finitely many stochastic-ODEs governing the particle system (SSA), to the limiting dynamics for the probability measure, $\mu_t(\mathbf{x}, m)$, governed by the limiting PDE (2.4); and second, the large-time behavior of the latter, summarized in Theorem 2.4, implies that in this mean-field limit, the PDE in the long time returns the provisional minimum as the global minimum. Combining these results we conclude in Theorem 2.5 the claimed convergence $\liminf_{t\to\infty} \lim_{N\to\infty} \overline{F}_t^N = F_*$ with a precise convergence rate of polynomial type. Further, in section 3 we derive the corresponding macroscopic description of the system.

1.2. The swarm-based optimization with adjusted annealing rate. The exchange of mass encoded in (SSA) assigns smaller masses to agents with higher values, while the decreasing property of $\sigma(m)$ ensures that these agents experience larger random perturbations. As a result, lighter agents serve as explorers, while heavier agents serve as leaders. As noted in [21], upon normalization of the total mass, $\sum_j m_t^j \equiv \sum_j m_0^j = 1$, masses can be interpreted as "probabilities of finding a global minimum"; heavier agents have larger probability in doing so. The computation proceeds by discretization of (SSA) using Euler–Maruyama's formula with a proper time step $h > 0$,

$$(1.2) \qquad \mathbf{x}_{n+1}^j = \mathbf{x}_n^j - h\nabla F(\mathbf{x}_n^j) + \sqrt{2h\sigma\left(m_n^j\right)}\boldsymbol{\xi}^j$$

for $j = 1, 2, \ldots, N$ and $\boldsymbol{\xi}^j \sim \mathcal{N}(0, \mathbf{I}_d)$ independent and identically distributed (i.i.d.). for all j. The detailed algorithm is summarized in Algorithm 1.1. The mass is conserved throughout the algorithm. According to line 6, the update gives $\sum_{j=1}^N m_{n+1}^j = \sum_{j=1}^N m_n^j - h\sum_{j=1}^N m_n^j\left(F(\mathbf{x}_n^j) - \overline{F}_n\right) = \sum_{j=1}^N m_n^j$, where the second summation term becomes zero owing to the definition of $\overline{F}_n$ shown in line 3 in the algorithm.

The performance of the algorithm is presented in section 6. This follows the analytical study in sections 2 and 5 in which we state and, respectively, prove the main analytical results of the paper, regarding the large-crowd, large-time dynamics associated with (SSA).

Algorithm 1.1 SSA.

1: **Input**: Initial distribution $\mu_0(x,m)$, h as step size, and n_T as # of iterations

2: $\{(\mathbf{x}_0^j, m_0^j)\}_{j=1}^N$ are i.i.d. drawn from μ_0 for $j = 1, 2, \ldots, N$

3: $\overline{F}_0 \leftarrow \frac{\sum_{j=1}^N m_0^j F(\mathbf{x}_0^j)}{\sum_{j=1}^N m_0^j}$ % $\overline{F}_n \equiv \overline{F}_n^N = \frac{\sum_{j=1}^N m_n^j F(\mathbf{x}_n^j)}{\sum_{j=1}^N m_n^j}$ the provisional minimum at t^n

4: **for** $n = 0, 1, 2, \ldots, n_T - 1$ **do**

5:     **for** $j = 1, 2, \ldots, N$ **do**

6:         $m_{n+1}^j \leftarrow m_n^j - m_n^j h\big(F(\mathbf{x}_n^j) - \overline{F}_n\big)$ for $j = 1, 2, \ldots, N$

7:     **end for**

8:     Generate N samples i.i.d. such that $\boldsymbol{\xi}^j \sim \mathcal{N}(0, \mathbf{I}_d)$

9:     $\mathbf{x}_{n+1}^j \leftarrow \mathbf{x}_n^j - h\nabla F(\mathbf{x}_n^j) + \sqrt{2h\sigma(m_n^j)}\boldsymbol{\xi}^j$ for $j = 1, 2, \ldots, N$

10:     $\overline{F}_{n+1} \leftarrow \sum_{j=1}^N m_{n+1}^j F(\mathbf{x}_{n+1}^j) / \sum_{j=1}^N m_{n+1}^j$

11: **end for**

12: $j_{\text{opt}} \leftarrow \operatorname{argmin}_j F(\mathbf{x}_{n+1}^j)$

13: **Output**: $\mathbf{x}_{n_T}^{j_{\text{opt}}}$ and $\overline{F}_{n_T}$

1.3. Related work. There are several well-known nonconvex optimization algorithms which combine swarm-based and/or stochastic effects. We mention biologically inspired methods of ant colony optimization [23], artificial bee colony optimization [16], and firefly optimization [31], and physically inspired methods of wind-driven optimization, [1]. We also mention particle swarm optimization methods which explore the state space with randomized drifts toward the best global position, [17, 26] and its stochastic version in [14].

The stochastic part of our method is motivated by SA driven by stochastic noise that is "cooled down" as time evolves, [15, 24]. The key feature in one-particle SA dynamics is the protocol for cooling down, i.e., setting a properly tuned decreasing-in-time annealing rate $\sigma_t \propto c/\sqrt{\log t}$ [13, 12, 11, 19, 7]. This should be compared with our swarm-based approach, in which the annealing rate of each agent is dictated by a decreasing function of its mass,[2] $\sigma_t \mapsto \sigma(m_t^j)$; essentially, we let the swarm decide how to cool down the agents as they accumulate more mass.

The idea of gaining from the stochastic interaction of more than just one agent can be found in the more recent work [6]. Here Langevin diffusion swaps between two agents—a "global explorer" and a "local explorer," with two annealing rates which correspond to high and low temperatures. The dynamics takes place in $\mathbb{R}^d \times \mathbb{R}^d$. In our method, there is a crowd of N agents which evolve their mass-dependent annealing rates in $(\mathbb{R}^d \times \mathbb{R}_+)^{\otimes N}$.

In the present context, it is instructive to calibrate our method with the consensus based optimization (CBO), [25, 2, 30, 3, 4, 8]. The CBO is governed by a swarm-based dynamics of SDEs,

(1.3)

$$d\mathbf{x}_t^j = -\lambda\left(\mathbf{x}_t^j - \overline{F}_t^\lambda\right)dt + \sigma\left|\mathbf{x}_t^j - \overline{F}_t^\lambda\right|dW_t^j, \qquad \overline{F}_t^\lambda := \sum_{j=1}^N \mathbf{x}_t^j\left(\frac{\exp(-\lambda F(\mathbf{x}_t^j))}{\sum_{i=1}^N \exp(-\lambda F(\mathbf{x}_t^i))}\right),$$

---

[2]In fact, we allow a larger class of mass scaling, covered in Assumptions 2.1 and 2.3.

which are steering towards the weighted minimum $\overline{F}_t^\lambda$. Indeed, the weighted min $\overline{F}_t^\lambda$ can be viewed as the provisional minimum driving the CBO (1.3), in view of the Laplace principle,

$$\lim_{\lambda\to\infty}\left(-\frac{1}{\lambda}\log\left(\int\omega_F^\lambda(\mathbf{x})\mathrm{d}\rho\right)\right)=\min_{\mathrm{supp}(\rho)}F,\qquad \omega_F^\lambda(\mathbf{x}):=\exp(-\lambda F(\mathbf{x})).$$

It is here that we observe the main difference between our swarm-based optimization with adjusted annealing rate (SSA) and the CBO (1.3): while the latter requires a user tuning of parameter λ in $\overline{F}^\lambda$, the swarm-based framework of [21] provides us with an *adaptive* mechanism which dynamically adjusts the provisional minimum based on the mass distribution in different parts of the landscape, $\overline{F}_t^N$, and the corresponding mean-field $\overline{\mathscr{F}}_t^\mu$, outlined in (2.4) below.

2. Statement of main results.

2.1. From empirical distribution to mean field. We define the *empirical distribution* μ_t^N, which records the *ensemble* distribution of $\{\mathbf{x}_t^j, m_t^j\}_{j=1}^N$ that obey (SSA), as a measure on the feature space $(\mathbf{x}, m)$:

$$(2.1)\qquad \mu_t^N(\mathbf{x}, m)=\frac{1}{N}\sum_{j=1}^N\delta_{\mathbf{x}_t^j}(\mathbf{x})\otimes\delta_{m_t^j}(m).$$

Recall that the provisional minimum is given by the weighted average in (1.1), which is expressed as

$$(2.2)\qquad \overline{F}_t^N=\frac{\displaystyle\sum_{j=1}^N m_t^j F(\mathbf{x}_t^j)}{\displaystyle\sum_{j=1}^N m_t^j}=\frac{\displaystyle\iint mF(\mathbf{x})\mathrm{d}\mu_t^N(\mathbf{x}, m)}{\displaystyle\iint m\mathrm{d}\mu_t^N(\mathbf{x}, m)}.$$

To formally derive the mean-field limit, we use a smooth test function $\phi(\mathbf{x}, m)$ that is compactly supported on $\mathbb{R}^d\times\mathbb{R}$. By definition,

$$\iint\phi(\mathbf{x}, m)\mu_t^N(\mathbf{x}, m)\mathrm{d}x\mathrm{d}m=\frac{1}{N}\sum_{j=1}^N\phi(\mathbf{x}_t^j, m_t^j).$$

Differentiating, recalling Itô's formula, and using (SSA) yield

$$\frac{1}{N}\sum_{j=1}^N\mathbb{E}\left[\mathrm{d}\phi(\mathbf{x}_t^j, m_t^j)\right]$$

$$=\frac{1}{N}\sum_{j=1}^N\mathbb{E}\left[\nabla_{\mathbf{x}}\phi(\mathbf{x}_t^j, m_t^j)\cdot\mathrm{d}\mathbf{x}_t^j+\frac{1}{2}\langle D_{\mathbf{x}}^2\phi, 2\sigma(m_t^j)\rangle_{\mathrm{Tr}}\mathrm{d}t+\partial_m\phi(\mathbf{x}_t^j, m_t^j)\mathrm{d}m_t^j\right]$$

$$=\frac{1}{N}\sum_{j=1}^N\mathbb{E}\underbrace{\left[\nabla_{\mathbf{x}}\phi\cdot\left(-\nabla F(\mathbf{x}_t^j)\mathrm{d}t+\sqrt{2\sigma(m_t^j)}\mathrm{d}W_t^j\right)\right]}_{\text{Term I}}+\frac{1}{N}\sum_{j=1}^N\mathbb{E}\underbrace{\left[\langle D_{\mathbf{x}}^2\phi, \sigma(m_t^j)\rangle_{\mathrm{Tr}}\mathrm{d}t\right]}_{\text{Term II}}$$

$$+\frac{1}{N}\sum_{j=1}^N\mathbb{E}\underbrace{\left[-\partial_m\phi(\mathbf{x}_t^j, m_t^j)(F(\mathbf{x}_t^j)-\overline{F}_t^N)m_t^j\mathrm{d}t\right]}_{\text{Term III}},$$

SWARM-BASED GD MEETS SIMULATED ANNEALING

where $D_{\mathbf{x}}^2\phi$ is the Hessian. The first term we have is, $\mathrm{I} = \mathrm{I}_1 + \mathrm{I}_2$, where

$$\mathrm{I}_1 := \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[\nabla_{\mathbf{x}}\phi(\mathbf{x}_t^j, m_t^j)\cdot -\nabla F(\mathbf{x}_t^j)\right] = -\iint \nabla_{\mathbf{x}}\phi(\mathbf{x}, m)\cdot\nabla F(\mathbf{x})\mu_t^N(\mathbf{x}, m)\mathrm{d}\mathbf{x}\mathrm{d}m$$

$$= \iint \phi(\mathbf{x}, m)\nabla_{\mathbf{x}}\cdot(\mu_t^N\nabla F(\mathbf{x}))\mathrm{d}\mathbf{x}\mathrm{d}m,$$

and $\mathrm{I}_2 := \mathbb{E}[\mathrm{d}W_t^j] = 0$. For the second term, we have

$$\mathrm{II} := \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[\langle D_{\mathbf{x}}^2\phi(\mathbf{x}_t^j, m_t^j), \sigma(m_t^j)\rangle_{\mathrm{Tr}}\mathrm{d}t] = \iint \Delta_{\mathbf{x}}\phi(\mathbf{x}, m)\sigma(m)\mu_t^N(\mathbf{x}, m)\mathrm{d}\mathbf{x}\mathrm{d}m$$

$$= \iint \phi(\mathbf{x}, m)\sigma(m)\Delta_{\mathbf{x}}\mu_t^N(\mathbf{x}, m)\mathrm{d}\mathbf{x}\mathrm{d}m,$$

and finally, for the third term

$$\mathrm{III} := \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\left[-\partial_m\phi(\mathbf{x}_t^j m_t^j)m_t^j(F(\mathbf{x}_t^j) - \overline{F}_t^N)\right]$$

$$= -\iint \partial_m\phi(\mathbf{x}, m)(m(F(\mathbf{x}) - \overline{F}_t^N))\mu_t^N(\mathbf{x}, m)\mathrm{d}\mathbf{x}\mathrm{d}m$$

$$= \iint \phi(\mathbf{x}, m)\partial_m(m\mu_t^N(\mathbf{x}, m))(F(\mathbf{x}) - \overline{F}_t^N)\mathrm{d}\mathbf{x}\mathrm{d}m.$$

Adding all these terms together, we conclude that the dynamics of the empirical distribution, μ_t^N, is governed by the Vlasov equation

$$(2.3) \qquad \partial_t\mu_t^N = \nabla_{\mathbf{x}}\cdot(\mu_t^N\nabla F) + (F(\mathbf{x}) - \overline{F}_t^N)\partial_m(m\mu_t^N) + \sigma(m)\Delta_{\mathbf{x}}\mu_t^N.$$

This formally provides the mean-field equation. In the mean-field limit, $\mu = \mu_t(\mathbf{x}, m)$ is governed by

$$(2.4a) \qquad \partial_t\mu = \nabla_{\mathbf{x}}\cdot(\mu\nabla F) + (F(\mathbf{x}) - \overline{\mathscr{F}}_t^\mu)\partial_m(m\mu) + \sigma(m)\Delta_{\mathbf{x}}\mu,$$

driven by the corresponding provisional minimum

$$(2.4b) \qquad \overline{\mathscr{F}}_t^\mu := \frac{\mathbb{E}_\mu[mF(\mathbf{x})]}{\mathbb{E}_\mu[m]} = \frac{\iint mF(\mathbf{x})\mathrm{d}\mu_t(\mathbf{x}, m)}{\iint m\mathrm{d}\mu_t(\mathbf{x}, m)}.$$

Observe that $\overline{\mathscr{F}}_t^\mu$ denotes the provisional minimum associated with the mean-field μ, which is to be distinguished from the discrete provisional minimum in (2.2), denoted $\overline{F}_t^N$. We note that, similarly to the discrete mass conservation in (SSA), $\frac{1}{N}\sum_j m_t^j \equiv 1$, the total mass in (2.4a) is conserved and equals one:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int m\mathrm{d}\mu_t(\mathbf{x}, m) = 0, \quad \int m\mathrm{d}\mu_t(\mathbf{x}, m) = 1.$$

Our first result quantifies the convergence of the empirical distribution, μ_t^N, to its limiting mean field μ_t. To this end, we make the following assumption.

Assumption 2.1.
- Lipschitz bound—F, ∇F and $\sqrt{\sigma(\cdot)}$ are all Lipschitz continuous functions with the Lipschitz constant denoted by L.
- Boundedness—There exists $D > 0$ such that $|F| + \sqrt{\sigma} \leqslant D$. Without loss of generality, we set $D \geqslant 1$ and can further assume $F \geqslant 0$.
- Bounded support uniformly in m—there exists $M_{\max} < \infty$ such that

$$\mathrm{Supp}_{\mathbf{x}}(\mu_0(\mathbf{x}, m)) \subset [0, M_{\max}] \qquad \forall m \in (0, \infty).$$

The next theorem shows that the empirical distribution μ_t^N and the mean field μ stay close. The proof is postponed to section 5.1.

THEOREM 2.2 (mean-field limit). *Assume that Assumption* 2.1 *holds. Let* $\mu_t = \mu_t(\mathbf{x}, m)$ *be the mean-field solution of* (2.4a) *and let* $\mu_t^N = \frac{1}{N} \sum_{j=1}^N \delta_{\mathbf{x}_t^j}(\mathbf{x}) \otimes \delta_{m_t^j}(m)$ *be the empirical distribution associated with the ensemble of swarm-based solutions* (SSA) *subject to compatible initial data so that* $\{\mathbf{x}_0^j, m_0^j\}$ *are i.i.d. samples of* $\mu_{t=0}(\mathbf{x}, m)$. *Then* μ_t *and* μ_t^N *are close in the Wasserstein sense:*

$$(2.5) \qquad W_2\left(\mu_t, \mu_t^N\right) \to 0 \quad \text{in probability as } N \to \infty,$$

and the corresponding provisional minimum (2.2), $\overline{F}_t^N$, *converges to* $\overline{\mathscr{F}}_t^\mu$ *in* (2.4b) *with the law of large numbers rate: there exists a constant* $C_t = C_t(\mu_0, F, d) > 0$ *independent of* N *such that*

$$(2.6) \qquad \mathbb{E}\left[\left|\overline{F}_t^N - \overline{\mathscr{F}}_t^\mu\right|\right] < \frac{C_t}{\sqrt{N}}.$$

The bound (2.6) provides the crucial convergence bound, which translates the provisional minimum of the many-agent system to that of the limiting mean-field equation.

2.2. Large-time convergence—from mean field to global minimum. Upon translating the convergence of $\overline{F}_t^N$ to that of $\overline{\mathscr{F}}_t^\mu$, we now switch gear to study the convergence of the mean-field limiting PDE. To this end, we need another set of assumptions for F, σ, and μ_0.

Assumption 2.3. We assume that the following conditions hold true:
- $\sigma(m)$ has a bounded support: there exists an $m_c > 0$ such that $\sigma(m) = 0$ when $m > m_c$;
- $(m_c, \infty) \cap \mathrm{supp}_m(\mu_0) \neq \emptyset$ and $M(0) = \iint m \mathrm{d}\mu_0(\mathbf{x}, m) = 1$;
- F admits a unique global minimum $\mathbf{x}_* := \mathrm{argmin}_{\mathbf{x} \in \Omega} F(\mathbf{x})$ and, thus, $\nabla F(\mathbf{x}_*) = 0$.

The next theorem shows that the provisional minimum of the mean field, $\overline{\mathscr{F}}_t^\mu$, is converging towards the global minimum of F. The proof is postponed to section 5.2 below.

THEOREM 2.4 (large-time behavior). *Assume that Assumptions* 2.1, 2.3 *hold. Let* μ *be the mean field with provisional minimum* $\overline{\mathscr{F}}_t^\mu$. *We have*

$$(2.7) \qquad \liminf_{t \to \infty} \overline{\mathscr{F}}_t^\mu = \min_{\mathbf{x}} F(\mathbf{x}) = F_*.$$

Specifically,—for any $\epsilon > 0$, *we let* $\mathcal{N}_\epsilon$ *denote the following* ϵ *neighborhood:*

$$(2.8) \qquad \mathcal{N}_\epsilon := \{\mathbf{x} \mid F(\mathbf{x}) < F_* + \epsilon\}.$$

Then, there exists a time $t_\epsilon(m_c) < \infty$ defined by[3]

$$(2.9) \qquad t_\epsilon(m_c) = \frac{2}{\epsilon} \log\left(\left(\sup_{t\in[0,1]} \int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_t(\mathbf{x},m) \right)^{-1} \right) + 1,$$

and $t \leqslant t_\epsilon(m_c)$ such that

$$(2.10) \qquad \overline{\mathscr{F}}_t^\mu < F_* + \epsilon.$$

Observe that even if $\mathcal{N}_\epsilon \cap \mathrm{supp}_\mathbf{x}\mu_0(\cdot,m) = \emptyset$, then the presence of the Brownian motion implies that there exists a time, $t \in (0,1]$, such that $\mathcal{N}_\epsilon \cap \mathrm{supp}_\mathbf{x}\mu_t(\cdot,m) \neq \emptyset$ and, consequently, (2.9) always yields a finite number threshold, t_ϵ. The combination of (2.9)–(2.10) together claims that within roughly $t_\epsilon \approx \frac{1}{\epsilon}$ time, the global optimal value can be found within ϵ accuracy, implying an $\mathcal{O}(1/\epsilon)$ convergence. However, the constant dependence heavily relies on the initial data, and we cannot claim the optimality of the :convergence rate:' our numerical findings reported in section 6 indicate the possibility of a much better convergence rate in the sense that $t_\epsilon = \log(1/\epsilon)$ is already sufficient to ensure (2.10). See Figure 6.4.

Combining Theorems 2.2 and 2.4 we conclude the following main result of this paper, namely, the convergence of the empirical distribution to the mean field.

THEOREM 2.5. *Assume that Assumptions* 2.1, 2.3 *hold[4]. Let $\mu_t^N(\mathbf{x},m)$ be the empirical distribution associated with an ensemble of solutions of the swarm-based optimization with adjusted annealing rate* (SSA), $(\mathbf{x}_t^j, m_t^j)$, *subject to the initial data drawn i.i.d. from $\mu_{t=0}$. Fix an arbitrary $\epsilon > 0$. Then, for a large enough crowd spelled out in* (2.6), *depending on time, $t_{\epsilon/2}$ defined in* (2.9),

$$N > \frac{4C_{t_{\epsilon/2}}^2}{\epsilon^2} \quad \text{where} \quad t_{\epsilon/2} < \frac{4}{\epsilon} \log\left(\left(\sup_{t\in[0,1]} \int_{\mathcal{N}_{\epsilon/2}} \int_{m_c}^\infty m \mathrm{d}\mu_t(\mathbf{x},m) \right)^{-1} \right) + 1,$$

so that there is a $t < t_{\epsilon/2}$ and the provisional minimum $\overline{F}_t^N$ is within ϵ of $\min_\mathbf{x} F(\mathbf{x})$,

$$\mathbb{E}\left[\left| \overline{F}_t^N - \min_\mathbf{x} F(\mathbf{x}) \right| \right] < \epsilon.$$

Proof of Theorem 2.5. By Theorem 2.4, there is a $t < t_{\epsilon/2}$ so that

$$\left| \overline{\mathscr{F}}_t^\mu - \min_\mathbf{x} F(\mathbf{x}) \right| < \frac{\epsilon}{2},$$

and according to Theorem 2.2, there exists a constant, C_t such that

$$\mathbb{E}\left[\left| \overline{F}_t^N - \overline{\mathscr{F}}_t^\mu \right| \right] < \frac{C_t}{\sqrt{N}}.$$

Hence, for large enough N such that $\frac{C_t}{\sqrt{N}} < \frac{\epsilon}{2}$ or, equivalently, $N > \frac{4C_{t_{\epsilon/2}}^2}{\epsilon^2}$ (note $C_{t_{\epsilon/2}} > C_t$), we conclude

---

[3]Note that $\int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_t(\mathbf{x},m) \leqslant \iint m \mathrm{d}\mu_t(\mathbf{x},m) = 1$ for mass conservation. The term in the logarithm is always bigger than 1, so that the expression on the right-hand side of (2.9) quantifies positive time.

[4]Assumption 2.3 requires the uniqueness of the global minimum. This assumption can be replaced by finite many global minima $\{\mathbf{x}_k\}_{k=1}^K$ with $F(\mathbf{x}_k) = F_*$ for all k. The details of the analysis need to be revised accordingly. In the proof (section 5.2), we point out the specific location to revise to accommodate the situation where multiple global minima coexist.

$$\mathbb{E}\left[\left|\overline{F}_t^N - \min_{\mathbf{x}} F(\mathbf{x})\right|\right] \leqslant \mathbb{E}\left[\left|\overline{F}_t^N - \overline{\mathscr{F}}_t^\mu\right|\right] + \left|\overline{\mathscr{F}}_t^\mu - \min_{\mathbf{x}} F(\mathbf{x})\right| \leqslant \epsilon. \qquad \square$$

3. From mean field to macroscopic description. We shall briefly comment on the macroscopic description of the swarm-based optimization with adjusted annealing rate, furnished in terms of the first two moments—the density, $\rho : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}_+$, and momentum, $\rho\mathsf{M} : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}$ (observe that time dependence is denoted as usual for the density $\rho = \rho(t, \mathbf{x})$, macroscopic mass, $\mathsf{M} = \mathsf{M}(t, \mathbf{x})$, etc.),

$$\rho(t, \mathbf{x}) := \int \mu_t(\mathbf{x}, m)\mathrm{d}m, \qquad \rho\mathsf{M}(t, \mathbf{x}) := \int m\mu_t(\mathbf{x}, m)\mathrm{d}m.$$

Taking the first two moments of the mean field (2.4), we find that

$$(3.1) \quad \begin{cases} \rho_t - \nabla_{\mathbf{x}} \cdot (\rho\nabla F) = \Delta_{\mathbf{x}} \int \sigma(m)\mu_t(\mathbf{x}, m)\mathrm{d}m, \\[2mm] (\rho\mathsf{M})_t - \nabla_{\mathbf{x}} \cdot (\rho\mathsf{M}\nabla F) = \big(\overline{F}(t) - F(\mathbf{x})\big)\rho\mathsf{M} + \Delta_{\mathbf{x}} \int m\sigma(m)\mu_t(\mathbf{x}, m)\mathrm{d}m. \end{cases}$$

We normalize the total mass and momentum, $\int \rho(t, \mathbf{x})\mathrm{d}\mathbf{x} \equiv 1$ and $\int \rho\mathsf{M}(t, \mathbf{x})\mathrm{d}\mathbf{x} \equiv 1$, which leaves us with a provisional minimum, $\overline{F}(t) = \overline{\mathscr{F}}_t^\mu$, given by

$$\overline{F}(t) = \int F(\mathbf{x})\rho\mathsf{M}(t, \mathbf{x})\mathrm{d}\mathbf{x}.$$

To unlock a closed form of the diffusion on the right, we assume the simplest closure based on a monokinetic pseudo-Maxwellian, $\mu_t(\mathbf{x}, m) = \rho(t, \mathbf{x})\delta(m - \mathsf{M}(t, \mathbf{x}))$. This yields

$$(3.2\mathrm{a}) \qquad \rho_t - \nabla_{\mathbf{x}} \cdot (\rho\nabla F) = \Delta_{\mathbf{x}}(\sigma(t, \mathbf{x})\rho(t, \mathbf{x})), \qquad \sigma(t, \mathbf{x}) := \sigma(\mathsf{M}(t, \mathbf{x})),$$

$$(3.2\mathrm{b}) \qquad (\rho\mathsf{M})_t - \nabla_{\mathbf{x}} \cdot (\rho\mathsf{M}\nabla F) = \big(\overline{F}(t) - F(\mathbf{x})\big)\rho\mathsf{M} + \Delta_{\mathbf{x}}\big(\sigma(t, \mathbf{x})\rho\mathsf{M}(t, \mathbf{x})\big).$$

The momentum (3.2b) can be converted into a drift-diffusion equation for the velocity M,

$$(3.3\mathrm{a}) \qquad \rho_t - \nabla_{\mathbf{x}} \cdot (\rho\nabla F) = \Delta_{\mathbf{x}}(\sigma(t, \mathbf{x})\rho(t, \mathbf{x})),$$

$$\mathsf{M}_t - \nabla F \cdot \nabla_{\mathbf{x}}\mathsf{M} = \big(\overline{F}(t) - F(\mathbf{x})\big)\mathsf{M} + \frac{1}{\rho}\Delta_{\mathbf{x}}\big(\sigma(t, \mathbf{x})\rho\mathsf{M}\big)$$

$$(3.3\mathrm{b}) \qquad\qquad\qquad - \frac{1}{\rho}\Delta_{\mathbf{x}}\big(\sigma(t, \mathbf{x})\rho\big)\mathsf{M}, \quad \mathbf{x} \in \mathrm{supp}\{\rho(t, \cdot)\}.$$

We shall not dwell on a detailed study of (3.3), but note the descent estimate

$$\frac{\mathrm{d}}{\mathrm{d}t} \int (F(\mathbf{x}) - F_*)\rho(t, \mathbf{x})\mathrm{d}\mathbf{x} = - \int |\nabla F(\mathbf{x})|^2 \rho(t, \mathbf{x})\mathrm{d}\mathbf{x} + \int \Delta_{\mathbf{x}}F(\mathbf{x})\sigma(t, \mathbf{x})\rho(t, \mathbf{x})\mathrm{d}\mathbf{x}$$

with *expected* long time behavior $\rho(t, \mathbf{x}) \overset{t\to\infty}{\longrightarrow} \delta(\mathbf{x} - \mathbf{x}_*)$ and $\mathsf{M}(t, \mathbf{x}) \overset{t\to\infty}{\longrightarrow} \mathbb{1}(\mathbf{x}_*)$. Observe that $\sigma(t, \mathbf{x})$ should vanish as $\mathbf{x} \to \mathbf{x}_*$.

4. Pitfalls. We examine the swarm-based optimization with adjusted annealing rate in two extreme cases–when $\sigma(m) \equiv 0$ and $\sigma(m) \equiv 1$. The former case turns off stochasticity and amounts to a deterministic system of N independent gradient descending agents; we show that their convergence fails in a probability sense. The latter turns on stochasticity uniformly for all agents, which amounts to a stochastic

system of N independent gradient descending agents; we show that their convergence fails in an accuracy sense.

We note that in both cases, the resulting systems are passive systems in the sense that the weights of agents do *not* affect their trajectories. As a consequence, agents roam the landscape with equal randomness with no distinction between leaders and explorers. The convergence of the provisional minimum $\overline{\mathscr{F}}_t^\mu$ is naturally worse than in the communicating swarm-based dynamics.

4.1. Swarming with no communication. What is the effect of Brownian motion in our swarm-based optimization with adjusted annealing rate? If we turn off the amplitude, $\sigma(m) \equiv 0$, then (SSA) is reduced to a deterministic system of *noninteracting* agents with no Brownian motion (and again we switch the notations of time dependence in deterministic quantities, $\mathbf{x}^j(t), m^j(t), \overline{F}^N(t)$, etc.):

$$(4.1) \qquad \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}^j(t) = -\nabla F(\mathbf{x}^j(t)), \\[2mm] \dfrac{\mathrm{d}}{\mathrm{d}t}m^j(t) = -m^j(t)\big(F(\mathbf{x}^j(t)) - \overline{F}^N(t)\big), \end{cases} \qquad j = 1, \ldots, N.$$

We still assume random initial configurations: the initial drawing is independent, with $\{\mathbf{x}^j(0)\}_{j=1}^N$ drawn from an initial distribution ρ_0, and $m^j(0) = \frac{1}{N}$. Since the total mass is conserved, we can let $M(t) = 1$, so that the provisional minimum is reduced to the usual average $\overline{F}^N(t) = \sum_{j=1}^N m^j(t)F(\mathbf{x}^j(t))$. The main feature *missing* in (4.1) is communication. Different agents at different positions proceed along gradient descent, independently of each other: since communication through a provisional minimum is missing, the dynamics lacks the collective engagement of the agents as a self-organized swarm.

The essential role of communication was already highlighted in the context of the SBGD method, [21]: once we decouple the step size and mass, setting a *uniform* step size in (SBGD), that is, $\alpha(m) \equiv 1$, then the success rate of the resulting noncommunicating swarm-based method decreases dramatically. The system (4.1) provides yet another realization of the essential role of communication. We note that since all particles are "descending," it is also straightforward to show that the collective mean objective $\overline{F}^N(t)$ is always decreasing: $\frac{\mathrm{d}}{\mathrm{d}t}\overline{F}^N(t) \leqslant 0$. See section A.1 for the proof. However, even though $\overline{F}^N(t)$ continues to decay in time, it does not guarantee the desired convergence to the global minimum, $\liminf_{t\to\infty} \overline{F}^N(t) = \min_{\mathbf{x}} F(\mathbf{x})$. This is due to the passive nature of the gradient descent of the agents. In particular, it is likely that all agents will become trapped at local minima or saddle points, especially when the number of particles is not sufficiently large. Turning on the communication term that adjusts the strength of the Brownian motion, as is done in (SSA), allows particles situated at local minima to experience increased randomness. This increased randomness facilitates their escape from local basins. With the communication and the randomness turned off, we have the following.

THEOREM 4.1 (lack of communication and failure in probability). *Assume F has a unique global minimum $\mathbf{x}_*$ inside an open set Ω so that $F_* = F(\mathbf{x}_*)$. Moreover, $\nabla F(\mathbf{x}) = 0$ on $\partial\Omega$. If we let $\{\mathbf{x}^j(0)\}$ drawn from ρ_0, $m^j(0) = \frac{1}{N}$, and evolve according to (4.1), then the success rate is bounded:*

$$(4.2) \qquad \mathbb{P}\left(\lim_{t\to\infty} \overline{F}^N(t) = F_*\right) \leqslant 1 - \left(1 - \int_\Omega \mathrm{d}\rho_0\right)^N.$$

In particular, if $\mathrm{supp}_{\mathbf{x}}(\rho_0) \cap \Omega = \emptyset$, *the success rate is* 0, *meaning* $\mathbb{P}(\lim_{t\to\infty} \overline{F}^N(t) = F_*) = 0$ *for all* N.

Proof. To prove (4.2), we first notice that without the Brownian motion, the deterministic system drives the samples to the local minima of the local basin in which the samples are located at $t = 0$, so the event $\{\lim_{t\to\infty} \overline{F}^N(t) = \min F\} \subset \{\exists \mathbf{x}^i(0) \in \Omega\}$. This implies

$$\mathbb{P}\left(\lim_{t\to\infty} \overline{F}^N(t) = \min F\right) \leqslant \mathbb{P}\left(\{\exists \mathbf{x}^j(0) \in \Omega\}\right) = 1 - \left(1 - \int_\Omega \mathrm{d}\rho_0\right)^N.$$

When $\mathrm{supp}_{\mathbf{x}}(\rho_0) \cap \Omega = \emptyset$, we have $\int_\Omega \mathrm{d}\rho_0 = 0$, so using this formula, we have for all N,

$$\mathbb{P}\left(\lim_{t\to\infty} \overline{F}^N(t) = \min F\right) = 0.$$

Since $\overline{F}^N(t)$ is a decreasing in t, the above equality leads to $\mathbb{P}(\lim_{t\to\infty} \overline{F}^N(t) > \min F) = 1$. □

This theorem explicitly spells out the "probability failure," so there is a nonzero probability that the mean objectives obtained by the particles cannot ever achieve the global optimizer.

4.2. Stochastic system. The limitation of (4.1) clearly originates from the fact that particle evolution closely resembles classical gradient descent, lacking a mechanism for domain exploration and avoiding entrapment in local minima. The other end of the spectrum is to introduce "too much" exploration. Indeed, setting $\sigma = 1$, we have

$$(4.3) \quad \begin{cases} \mathrm{d}\mathbf{x}_t^j = -\nabla F(\mathbf{x}_t^j)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t^j, \\ \mathrm{d}m_t^j = -m_t^j\left(F(\mathbf{x}_t^j) - \overline{F}_t^N\right)\mathrm{d}t, \end{cases} \quad j = 1, \ldots, N,$$

where W_t^j is the Brownian motion at time t associated with particle j. The only difference compared to (4.1) is the addition of a constant Brownian motion term. This Brownian motion term allows more effective exploration of the domain by preventing agents from becoming trapped at any local minima.

The trajectory of $\{\mathbf{x}_t^j\}$ resembles that of Langevin Monte Carlo, and under appropriate assumptions, it is well known that the distribution of $\mathbf{x}_t^j$ converges to $\rho_\infty \propto \exp(-F(\mathbf{x}))$ as $t \to \infty$. Incorporating mass exchange in (4.3) does not eliminate this issue. Carrying out the same mean-field analysis as was done in section 2.1, we find that the limiting PDE is

$$\partial_t \mu = \nabla_{\mathbf{x}} \cdot (\mu\nabla F) + \left(F(\mathbf{x}) - \overline{\mathscr{F}}_t^\mu\right)\partial_m(m\mu) + \Delta_{\mathbf{x}}\mu,$$

$$(4.4) \quad \overline{\mathscr{F}}_t^\mu = \frac{\iint mF(\mathbf{x})\mu_t(\mathbf{x},m)\mathrm{d}\mathbf{x}\mathrm{d}m}{\iint m\mu_t(\mathbf{x},m)\mathrm{d}\mathbf{x}\mathrm{d}m}.$$

Analyzing this mean-field PDE, we can show that there is always a positive gap between $\overline{\mathscr{F}}_t^\mu$ and F_*.

THEOREM 4.2. *Consider a function F satisfying the conditions outlined in Assumptions 2.1 and 2.3, and further assume it is uniformly strongly convex in a ball around the global minimum $\mathbf{x}_*$, meaning $\Delta_{\mathbf{x}} F \geqslant \xi d$ for $\xi > 0$ and $\mathbf{x} \in B_R(\mathbf{x}_*)$. Then there exists $\epsilon > 0$ (depending on $L, D, \xi, d,$ and R) such that the provisional minimum associated with the mean field μ in (4.4) satisfies*

$$(4.5) \qquad \overline{\mathscr{F}}_t^{\mu} \geqslant \min\{F_* + \epsilon, \overline{\mathscr{F}}_0^{\mu}\} \qquad \forall t > 0.$$

For brevity, the proof of this theorem is in Appendix A.2. This theorem explicitly spells out the "accuracy failure." The global optimizer cannot be achieved by the mean objective.

5. Proofs of the main results. In this section we prove Theorem 2.2 on the convergence to the mean-field limit, and Theorem 2.4 on the large-time behavior of the limiting equation.

5.1. Convergence to the mean-field limit. Theorem 2.2 states that the empirical distribution μ_t^N that assembles all samples that follow the coupled SDE system (SSA), is close to the limiting distribution μ that solves the PDE (2.4a).

There are a few different but related approaches for justifying mean-field limits, including the coupling method, the energy/entropy estimates, the utilization of the Bogoliubov–Born–Green–Kirkwood–Yvon hierarchies and running the tightness argument to confine a sequence of measures for the convergence; see references [22, 28]. Here we employ the coupling method, introduced in [28]. The coupling method places a stronger requirement on the field, but the machinery is rather simple to use, and it has the advantage of providing a precise convergence rate in terms of the number of particles, N.

To apply the coupling method, one usually designs an auxiliary system that is pushed forward by the underlying field. This auxiliary system serves as a bridge linking the discrete SDE system and the limiting mean-field PDE. On the one hand, it is driven by the underlying field and thus its ensemble closely resembles ρ, and, on the other hand, this system adopts the SDE structure and can be easily compared with the original self-consistent SDE system. In our context, we define such an auxiliary system as $\{\widetilde{\mathbf{x}}_t^j, \widetilde{m}_t^j\}_{j=1}^N$, that is governed by

$$(5.1) \qquad \begin{cases} \mathrm{d}\widetilde{\mathbf{x}}_t^j &= -\nabla F\left(\widetilde{\mathbf{x}}_t^j\right)\mathrm{d}t + \sqrt{2\sigma\left(\widetilde{m}_t^j\right)}\mathrm{d}W_t^j, \\ \mathrm{d}\widetilde{m}_t^j &= -\widetilde{m}_t^j(F\left(\widetilde{\mathbf{x}}_t^j\right) - \overline{\mathscr{F}}_t^{\mu})\mathrm{d}t \end{cases}$$

with the initial data set to be $\{\widetilde{\mathbf{x}}_0^j = \mathbf{x}_0^j, \widetilde{m}_0^j = m_0^j\}$, and $\overline{\mathscr{F}}^{\mu} = \mathbb{E}_{\mu}[mF(\mathbf{x})]/\mathbb{E}_{\mu}[m]$ was defined in (2.4b). For this system, we define the corresponding empirical distribution and the provisional minimum:

$$(5.2) \qquad \widetilde{\mu}_t^N = \frac{1}{N}\sum_i \delta_{\widetilde{\mathbf{x}}_t^j}(\mathbf{x}) \otimes \delta_{\widetilde{m}_t^j}(m) \quad \text{and} \quad \widetilde{F}_t^N = \frac{\frac{1}{N}\sum_j \widetilde{m}_t^j F\left(\widetilde{\mathbf{x}}_t^j\right)}{\mathbb{E}_{\mu}[m]}.$$

We note that this system has a clear similarity with the original SDE (SSA) but is passive, in the sense that the dynamics of $\widetilde{m}_t^j$ is driven by $\overline{\mathscr{F}}^{\mu}$ instead of $\widetilde{F}_t^N$. While the dynamics of $\widetilde{\mathbf{x}}_t$ agrees with that of $\mathbf{x}_t$, the mass changes call for different rates. In particular, in (SSA), we compare $F(\mathbf{x}_t)$ with $\overline{F}_t^N$ that is self-generated in (2.2) and, here, $F(\widetilde{\mathbf{x}}_t)$ is compared with $\overline{\mathscr{F}}^{\mu}$ defined from the underlying PDE system (2.4a).

As a consequence, this new $\{\widetilde{\mathbf{x}}_0^j = \mathbf{x}_0^j, \widetilde{m}_0^j = m_0^j\}$ system gets passively pushed forward by the underlying μ.

This design of the new auxiliary system prompts two lemmas.

LEMMA 5.1. *Suppose Assumption 2.1 holds. Let $\{\widetilde{\mathbf{x}}_t^j, \widetilde{m}_t^j\}_{j=1}^N$ be the system defined in (5.1) and let μ_t solve (2.4a). If the initial condition $\{\widetilde{\mathbf{x}}_0^j = \mathbf{x}_0^j, \widetilde{m}_0^j = m_0^j\}$ is drawn from $\mu_{t=0}$, then*
- μ_t and $\widetilde{\mu}_t^N$ are close in the Wasserstein sense:

$$W_2(\mu_t, \widetilde{\mu}_t^N) \to 0 \quad \text{in probability as } N \to \infty;$$

- *there exists a constant $C = C(t, d, \mu_0, F)$ independent of N such that $\widetilde{F}_t^N$ and $\overline{\mathscr{F}}_t^\mu$ are close,*

$$\mathbb{E}(|\widetilde{F}_t^N - \overline{\mathscr{F}}_t^\mu|) < \frac{C}{\sqrt{N}},$$

where $\widetilde{F}_t^N$ and $\overline{\mathscr{F}}_t^\mu$ are defined in (5.2) and (2.4b), respectively.

LEMMA 5.2. *Suppose Assumption 2.1 holds. Let $\{\widetilde{\mathbf{x}}_t^j, \widetilde{m}_t^j\}_{j=1}^N$ be the system defined in (5.1), and let $\{\mathbf{x}_t^j, m_t^j\}_{j=1}^N$ solve (SSA). Suppose the two systems have the same initial conditions, then*
- μ_t^N and $\widetilde{\mu}_t^N$ are close in the Wasserstein sense:

$$(5.3) \qquad W_2(\mu_t^N, \widetilde{\mu}_t^N) \to 0 \quad \text{in probability as } N \to \infty;$$

- *there exists a constant $C = C(t, d, \mu_0, F)$ independent of N such that $\widetilde{F}_t^N$ converges to $\overline{F}_t^N$ with the rate*

$$(5.4) \qquad \mathbb{E}(|\widetilde{F}_t^N - \overline{F}_t^N|) < \frac{C}{\sqrt{N}}.$$

Here $\widetilde{F}_t^N$ and $\overline{F}_t^N$ are defined in (5.2) and (2.2).

With these two lemmas, Theorem 2.2 is a straightforward corollary, calling triangle inequalities

$$W_2(\mu_t, \mu_t^N) \leqslant W_2(\mu_t, \widetilde{\mu}_t^N) + W_2(\widetilde{\mu}_t^N, \mu_t^N)$$

and

$$\mathbb{E}(|\overline{\mathscr{F}}_t^\mu - \overline{F}_t^N|) \leqslant \mathbb{E}(|\overline{\mathscr{F}}_t^\mu - \widetilde{F}_t^N|) + \mathbb{E}(|\widetilde{F}_t^N - \overline{F}_t^N|).$$

Here we omit the details and proceed to provide the proof of the two lemmas.

Proof of Lemma 5.1. Calling Theorem 1 in [9],

$$\mathbb{E}\left(W_2(\mu_t, \widetilde{\mu}_t^N)\right) = \mathcal{O}(N^{-\theta}) \to 0 \quad \text{as} \quad N \to \infty$$

with $\theta = \max\{1/2, 1/d\}$. This implies

$$W_2(\mu_t, \widetilde{\mu}_t^N) \to 0 \quad \text{in probability as } N \to \infty.$$

For weak convergence, we note that $(\widetilde{\mathbf{x}}, \widetilde{m})$ are i.i.d. samples from μ and thus

$$\mathbb{E}\left(\left|\frac{1}{N}\sum_{j=1}^N \widetilde{m}_t^j F\left(\widetilde{\mathbf{x}}_t^j\right) - \int mF(\mathbf{x})\mathrm{d}\mu_t(\mathbf{x}, m)\right|\right) < \frac{\mathrm{Var}_\mu^{1/2}(mF)}{\sqrt{N}} \leqslant \frac{C(t)M}{\sqrt{N}}. \qquad \square$$

The proof for Lemma 5.2 is the standard coupling method. This amounts to comparing the original SDE (SSA) with the auxiliary SDE (5.1). In particular, we note that the $\mathbf{x}$ component of the two systems are almost identical:

$$\mathbf{x}_t^j = -\nabla F(\mathbf{x}_t^j)\mathrm{d}t + \sqrt{2\sigma(m_t^j)}\mathrm{d}W_t^j \quad \text{versus} \quad \widetilde{\mathbf{x}}_t^j = -\nabla F(\widetilde{\mathbf{x}}_t^j)\mathrm{d}t + \sqrt{2\sigma(\widetilde{m}_t^j)}\mathrm{d}W_t^j.$$

Then we can subtract one from the other and trace the growth of the error. The same strategy is applied to tracking $m_t^j - \widetilde{m}_t^j$. The combination of the two pieces of error gives the control of the Wasserstein distance of $W_2(\mu_t^N, \widetilde{\mu}_t^N)$.

Proof of Lemma 5.2. Noting that $\mu_t^N = \frac{1}{N}\sum_j \delta_{\mathbf{x}_t^j}(\mathbf{x}) \otimes \delta_{m_t^j}(m)$ and $\widetilde{\mu}_t^N = \frac{1}{N}\sum_j \delta_{\widetilde{\mathbf{x}}_t^j}(\mathbf{x}) \otimes \delta_{\widetilde{m}_t^j}(m)$, then recalling the definition of the Wasserstein metric, we have

$$(5.5) \qquad W_2(\widetilde{\mu}_t^N, \mu_t^N)^2 \leqslant \left(\frac{1}{N}\sum_{j=1}^{N} \left[|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 + \left| m_t^j - \widetilde{m}_t^j \right|^2 \right] \right).$$

Define

$$\Omega = \left\{ \left| \frac{1}{N}\sum_{j=1}^{N} m_0^j - \mathbb{E}_{\mu_0}(m) \right| \leqslant \frac{1}{2} \right\}.$$

With Chebyshev's inequality and boundedness of m from Assumption 2.1, we obtain

$$\mathbb{P}(\Omega^c) \leqslant \frac{4\mathbb{E}_{\mu_0}(m^2)}{N} \leqslant \frac{C}{N} \quad \text{and thus, accordingly,} \quad \mathbb{P}(\Omega) \geqslant 1 - \frac{C}{N} \geqslant \frac{1}{2}.$$

Define

$$(5.6) \qquad G(t) = \mathbb{E}\left(\frac{1}{N}\sum_{j=1}^{N} \left[|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 + \left| m_t^j - \widetilde{m}_t^j \right|^2 \right] \Big| \Omega \right).$$

If we can show $G(t) \sim 1/N$, then conditioned on Ω, a set whose probability goes to 1, $W_2(\widetilde{\mu}_t^N, \mu_t^N)^2 \to 0$ as $N \to \infty$, proving (5.3).

Similarly, to estimate (5.4), we note

$$(5.7)$$
$$|\widetilde{F}_t^N - \overline{F}_t^N|$$
$$\leqslant \left| \frac{\frac{1}{N}\sum_{j=1}^{N} \left(\widetilde{m}_t^j F(\widetilde{\mathbf{x}}_t^j) - m_t^j F(\mathbf{x}_t^j) \right)}{\mathbb{E}_\mu[m]} \right| + \left| \left(\frac{1}{N}\sum_{j=1}^{N} m_t^j F(\mathbf{x}_t^j) \right) \left(\frac{1}{\mathbb{E}_\mu[m]} - \frac{1}{\frac{1}{N}\sum m_t^j} \right) \right|$$
$$\underset{(\mathrm{i})}{\leqslant} C \left[\left| \frac{1}{N}\sum_{j=1}^{N} \left(m_t^j - \widetilde{m}_t^j \right) F(\widetilde{\mathbf{x}}_t^j) + \frac{1}{N}\sum_{j=1}^{N} m_t^j \left(F(\mathbf{x}_t^j) - F(\widetilde{\mathbf{x}}_t^j) \right) \right| + \left| \frac{\frac{1}{N}\sum m_t^j - \mathbb{E}_\mu[m]}{\mathbb{E}_\mu[m]\left(\frac{1}{N}\sum m_t^j \right)} \right| \right]$$
$$\underset{(\mathrm{ii})}{=} C \left[\left| \frac{1}{N}\sum_{j=1}^{N} \left(m_t^j - \widetilde{m}_t^j \right) F(\widetilde{\mathbf{x}}_t^j) + \frac{1}{N}\sum_{j=1}^{N} m_t^j \left(F(\mathbf{x}_t^j) - F(\widetilde{\mathbf{x}}_t^j) \right) \right| \right.$$
$$\left. + \left| \frac{\frac{1}{N}\sum m_0^j - \mathbb{E}_{\mu_0}[m]}{\mathbb{E}_{\mu_0}[m]\left(\frac{1}{N}\sum m_0^j \right)} \right| \right]$$
$$\underset{(\mathrm{iii})}{\leqslant} C \left[\sqrt{\frac{1}{N}\sum_{j=1}^{N} \left[|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 + \left| m_t^j - \widetilde{m}_t^j \right|^2 \right]} + \left| \frac{1}{N}\sum m_0^j - \mathbb{E}_{\mu_0}[m] \right| \right].$$

Here in (i) we used the upper boundedness of both[5] m and F. In (ii), we apply the conservation of mass: $\sum m_t^j = \sum m_0^j$ and $\mathbb{E}_\mu[m] = \mathbb{E}_{\mu_0}[m]$. In (iii), we used the triangle inequality, the Cauchy–Schwarz inequality, and Lipschitz continuity of F. Consequently, the constant C depends on D, L, and $M_{\max}$ (from Assumption 2.1).

To insert it into (5.4), then

$$\mathbb{E}(|\widetilde{F}_t^N - \overline{F}_t^N|) \leqslant \mathbb{E}\left(|\widetilde{F}_t^N - \overline{F}_t^N||\Omega\right)\mathbb{P}(\Omega) + \mathbb{E}\left(|\widetilde{F}_t^N - \overline{F}_t^N||\Omega^c\right)\mathbb{P}(\Omega^c).$$

So combining with (5.7), we rewrite (5.4):

(5.8)
$$\mathbb{E}(|\widetilde{F}_t^N - \overline{F}_t^N|) \leqslant \mathbb{E}\left(|\widetilde{F}_t^N - \overline{F}_t^N||\Omega\right) + \frac{C}{N}$$

$$\leqslant C\,\mathbb{E}\left(\underbrace{\sqrt{\frac{1}{N}\sum_{j=1}^N\left[|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 + \left|m_t^j - \widetilde{m}_t^j\right|^2\right]}\,\bigg|\,\Omega}_{\leqslant \sqrt{G(t)}}\right) + C\,\underbrace{\mathbb{E}\left(\left|\frac{1}{N}\sum m_0^j - \mathbb{E}_{\mu_0}[m]\right|\bigg|\Omega\right)}_{\leqslant \frac{1}{\mathbb{P}(\Omega)}\mathbb{E}\left(\left|\frac{1}{N}\sum m_0^j - \mathbb{E}_{\mu_0}[m]\right|\right) \leqslant \frac{C}{\sqrt{N}}} + \frac{C}{N}.$$

Once again, if we can show $G(t) \sim 1/N$, (5.4) is complete.

The rest of the proof is dedicated to showing the bound for $G(t)$. To do so, we will trace the evolution of $|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2$ and $|m_t^j - \widetilde{m}_t^j|^2$, respectively.

- To trace $|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2$, we compare the $\mathbf{x}_t^j$-equation in (SSA) and the $\widetilde{\mathbf{x}}_t^j$-equation in (5.1) to have

$$g(\mathbf{x}) = |\mathbf{x}|^2, \quad \nabla_{\mathbf{x}}g(\mathbf{x}) = 2\mathbf{x}, \quad D_{\mathbf{x}}^2 g = 2\mathbb{I}.$$

Then $g(\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j) = |\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 = \langle \mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j, \mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j\rangle$, and with Itô's formula we get

$$\mathrm{d}g(\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j)$$

$$= \left(\langle\nabla g, -(\nabla F(\mathbf{x}_t^j) - \nabla F(\widetilde{\mathbf{x}}_t^j))\rangle + \mathrm{Tr}\left(\left(\sqrt{\sigma(m_t^j)} - \sqrt{\sigma(\widetilde{m}_t^j)}\right)^2 D_{\mathbf{x}}^2 g\right)\right)\mathrm{d}t$$

$$\quad + \left\langle\nabla g, \sqrt{2\sigma(m_t^j)} - \sqrt{2\sigma(\widetilde{m}_t^j)}\right\rangle\mathrm{d}W_t^j$$

$$= \left(-2\langle\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j, \nabla F(\mathbf{x}_t^j) - \nabla F(\widetilde{\mathbf{x}}_t^j)\rangle + 2d\left|\sqrt{\sigma(m_t^j)} - \sqrt{\sigma(\widetilde{m}_t^j)}\right|^2\right)\mathrm{d}t$$

$$\quad + 2\left\langle\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j, \sqrt{2\sigma(m_t^j)} - \sqrt{2\sigma(\widetilde{m}_t^j)}\right\rangle\mathrm{d}W_t^j.$$

Taking the expectation of the whole formula, the mean of the Brownian motion gets dropped out and we have

---

[5]First, we notice that, according to the boundedness in m condition in Assumption 2.1, the initial data for m_t^j are bounded by $M_{\max}$, meaning $|m_0^j| < M_{\max}$ for all j at the initial time. Given that $|F| \leqslant D$, then according to the m_t^j equation in (SSA), we call Grönwall's inequality to obtain that $|m_t^j| \leqslant M_{\max}e^{2Dt}$ for all time. The same argument applies to bound $\widetilde{m}_t^j$. We denote the upper bound c that depends on $M_{\max}$, D, and t.

$$(5.9) \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[g(\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j)|\Omega] = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2|\Omega]$$
$$\leqslant 2(L + L^2)\left(\mathbb{E}[|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2|\Omega] + d\mathbb{E}[|m_t^j - \widetilde{m}_t^j|^2|\Omega]\right),$$

where we used from Assumption 2.1 that both ∇F and $\sqrt{\sigma}$ are L-Lipschitz.
• To trace $|m_t^j - \widetilde{m}_t^j|^2$, we compare the m-equation in (SSA) and the m-equation in (5.1) to have

(5.10)
$$\frac{\mathrm{d}}{\mathrm{d}t}|m_t^j - \widetilde{m}_t^j|^2$$
$$= -2(m_t^j - \widetilde{m}_t^j)(m_t^j F(\mathbf{x}_t^j) - \widetilde{m}_t^j F(\widetilde{\mathbf{x}}_t^j))$$
$$+ 2(m_t^j - \widetilde{m}_t^j)\left(\frac{\frac{1}{N}\sum_{i=1}^N m_t^i F(\mathbf{x}_t^i)}{\frac{1}{N}\sum_{i=1}^N m_t^i} - \frac{\mathbb{E}_\mu[mF(\mathbf{x})]}{\mathbb{E}_\mu(m)}\right)$$
$$= -2(m_t^j - \widetilde{m}_t^j)^2 F(\mathbf{x}_t^j) - 2(m_t^j - \widetilde{m}_t^j)\widetilde{m}_t^j\left(F(\mathbf{x}_t^j) - F(\widetilde{\mathbf{x}}_t^j)\right)$$
$$+ 2(m_t^j - \widetilde{m}_t^j)\left(\frac{1}{N}\sum_{i=1}^N m_t^i F(\mathbf{x}_t^i) - \frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i)\right)/\left(\frac{1}{N}\sum_{i=1}^N m_t^i\right)$$
$$+ 2(m_t^j - \widetilde{m}_t^j)\left(\frac{\frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i)}{\frac{1}{N}\sum_{i=1}^N m_t^i} - \frac{\mathbb{E}_\mu[mF(\mathbf{x})]}{\mathbb{E}_\mu(m)}\right).$$

Reordering all terms, we have

(5.11)
$$\frac{\mathrm{d}}{\mathrm{d}t}|m_t^j - \widetilde{m}_t^j|^2 = \underbrace{-2(m_t^j - \widetilde{m}_t^j)^2 F(\mathbf{x}_t^j)}_{\text{(i)}} \underbrace{-2(m_t^j - \widetilde{m}_t^j)\widetilde{m}_t^j(F(\mathbf{x}_t^j) - F(\widetilde{\mathbf{x}}_t^j))}_{\text{(ii)}}$$
$$+ \underbrace{2(m_t^j - \widetilde{m}_t^j)\left(\frac{1}{N}\sum_{i=1}^N (m_t^i - \widetilde{m}_t^i)F(\mathbf{x}_t^i)\right)/\left(\frac{1}{N}\sum_{i=1}^N m_t^i\right)}_{\text{(iii)}}$$
$$+ \underbrace{2(m_t^j - \widetilde{m}_t^j)\left(\frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i(F(\mathbf{x}_t^i) - F(\widetilde{\mathbf{x}}_t^i))\right)/\left(\frac{1}{N}\sum_{i=1}^N m_t^i\right)}_{\text{(iv)}}$$
$$+ \underbrace{2(m_t^j - \widetilde{m}_t^j)\left(\frac{\frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i)}{\frac{1}{N}\sum_{i=1}^N m_t^i} - \frac{\mathbb{E}_\mu[mF(\mathbf{x})]}{\mathbb{E}_\mu(m)}\right)}_{\text{(v)}}.$$

We analyze each term separately:
  (i) According to Assumption 2.1, $F(\mathbf{x}_t^j) \geqslant 0$, so, $-2(m_t^j - \widetilde{m}_t^j)^2 F(\mathbf{x}_t^j) \leqslant 0$.
  (ii) Using Young's inequality,

$$-2(m_t^j - \widetilde{m}_t^j)\widetilde{m}_t^j\left(F(\mathbf{x}_t^j) - F\left(\widetilde{\mathbf{x}}_t^j\right)\right)$$
$$\leqslant \widetilde{m}_t^j\left(|m_t^j - \widetilde{m}_t^j|^2 + \left|F(\mathbf{x}_t^j) - F\left(\widetilde{\mathbf{x}}_t^j\right)\right|^2\right)$$
$$\leqslant c[|m_t^j - \widetilde{m}_t^j|^2 + L^2|\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2],$$

where c comes from the boundedness of m, seen in Footnote 5 and L is the Lipschitz constant from Assumption 2.1.

(iii) Because of the total mass preservation and the definition of Ω, we have

$$\frac{1}{N}\sum_{i=1}^{N} m_t^i = \frac{1}{N}\sum_{i=1}^{N} m_0^i \geqslant \frac{1}{2}\,.$$

Thus,

$$2(m_t^j - \widetilde{m}_t^j)\left(\frac{1}{N}\sum_{i=1}^{N}(m_t^i - \widetilde{m}_t^i)F(\mathbf{x}_t^i)\right)\bigg/\left(\frac{1}{N}\sum_{i=1}^{N} m_t^i\right)$$

$$\leqslant 2|m_t^j - \widetilde{m}_t^j|^2 + 2\left(\frac{1}{N}\sum_{i=1}^{N}(m_t^i - \widetilde{m}_t^i)F(\mathbf{x}_t^i)\right)^2$$

$$\leqslant D^2\left[|m_t^j - \widetilde{m}_t^j|^2 + \left(\frac{1}{N}\sum_{i=1}^{N}(m_t^i - \widetilde{m}_t^i)\right)^2\right]$$

$$\leqslant D^2\left[|m_t^j - \widetilde{m}_t^j|^2 + \frac{1}{N}\sum_{i=1}^{N}|m_t^i - \widetilde{m}_t^i|^2\right],$$

where we used Young's inequality in the first inequality, boundedness of F in the second (following Assumption 2.1), and the Hölder's inequality in the last.

(iv) Using the same argument as in (iii) we get

$$2(m_t^j - \widetilde{m}_t^j)\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{m}_t^i(F(\mathbf{x}_t^i) - F(\widetilde{\mathbf{x}}_t^i))\right)\bigg/\left(\frac{1}{N}\sum_{i=1}^{N} m_t^i\right)$$

$$\leqslant 2|m_t^j - \widetilde{m}_t^j|^2 + 2\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{m}_t^i(F(\mathbf{x}_t^i) - F(\widetilde{\mathbf{x}}_t^i))\right)^2$$

$$\leqslant c\left[|m_t^j - \widetilde{m}_t^j|^2 + \left(\frac{1}{N}\sum_{i=1}^{N}(F(\mathbf{x}_t^i) - F(\widetilde{\mathbf{x}}_t^i))\right)^2\right]$$

$$\leqslant c\left[|m_t^j - \widetilde{m}_t^j|^2 + \frac{1}{N}\sum_{i=1}^{N}|F(\mathbf{x}_t^i) - F(\widetilde{\mathbf{x}}_t^i)|^2\right]$$

$$\leqslant c\left[|m_t^j - \widetilde{m}_t^j|^2 + \frac{L^2}{N}\sum_{i=1}^{N}|\mathbf{x}_t^i - \widetilde{\mathbf{x}}_t^i|^2\right],$$

where we used Young's inequality in the first inequality, boundedness of $\widetilde{m}_t^j$ in the second (see Footnote 5), Hölder's in the third, and finally the Lipschitz continuity of F in the last inequality.

(v) Using the same argument,

$$2(m_t^j - \widetilde{m}_t^j)\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i)}{\frac{1}{N}\sum_{i=1}^{N} m_t^i} - \frac{\mathbb{E}_\mu[mF(\mathbf{x})]}{\mathbb{E}_\mu(m)}\right)$$

$$= 2(m_t^j - \widetilde{m}_t^j)\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i) - \mathbb{E}_\mu[mF(\mathbf{x})]\right)\bigg/\left(\frac{1}{N}\sum_{i=1}^{N} m_t^i\right)$$

$$+ 2(m_t^j - \widetilde{m}_t^i)\mathbb{E}_\mu[mF(\mathbf{x})]\left(\frac{1}{\frac{1}{N}\sum_{i=1}^N m_t^i} - \frac{1}{\mathbb{E}_\mu(m)}\right)$$

$$\leqslant 4|m_t^j - \widetilde{m}_t^j|^2 + 2\left|\frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i) - \mathbb{E}_\mu[mF(\mathbf{x})]\right|^2$$

$$+ 2D^2\left|\frac{1}{N}\sum_{i=1}^N m_0^i - \mathbb{E}_{\mu_0}(m)\right|^2,$$

where we use $\frac{1}{N}\sum_{i=1}^N m_t^i = \frac{1}{N}\sum_{i=1}^N m_0^i \geqslant \frac{1}{2}$, $\mathbb{E}_\mu(m) = \mathbb{E}_{\mu_0}(m)$, and the boundedness of F in the last inequality.

Collecting these terms together to feed into (5.11), and sum up against j, we have

(5.12)

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N |m_t^j - \widetilde{m}_t^j|^2\Big|\Omega\right]\right)$$

$$\leqslant \widetilde{C}\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N |\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 + \frac{1}{N}\sum_{j=1}^N |m_t^j - \widetilde{m}_t^j|^2\Big|\Omega\right]$$

$$+ \widetilde{C}\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i) - \mathbb{E}_\mu[mF(\mathbf{x})]\right|^2 + \left|\frac{1}{N}\sum_{i=1}^N m_0^i - \mathbb{E}_{\mu_0}(m)\right|^2\Big|\Omega\right],$$

where $\widetilde{C}$ depends on D, $M_{\max}$, t, and L.

Recall the definition of G in (5.6). We collect (5.9) and (5.12) to have

(5.13)

$$\frac{\mathrm{d}}{\mathrm{d}t}G(t) \leqslant C(t)\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N |\mathbf{x}_t^j - \widetilde{\mathbf{x}}_t^j|^2 + \frac{1}{N}\sum_{j=1}^N |m_t^j - \widetilde{m}_t^j|^2\Big|\Omega\right]$$

$$+ \underbrace{C(t)\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \widetilde{m}_t^i F(\widetilde{\mathbf{x}}_t^i) - \mathbb{E}_\mu[mF(\mathbf{x})]\right|^2 + \left|\frac{1}{N}\sum_{i=1}^N m_0^i - \mathbb{E}_{\mu_0}(m)\right|^2\Big|\Omega\right]}_{\text{(vi)}}$$

$$= C(t)\left(G(t) + \text{(vi)}\right),$$

where $C(t)$ depends on D, $M_{\max}$, t, L, and d. Note that term (vi) can be further controlled using the law of large numbers: for the first term, let $X = mF(\mathbf{x})$ be a random variable. Then, $\widetilde{m}_t^j F\left(\widetilde{\mathbf{x}}_t^j\right)$ is a realization of X and we have

$$\overline{X}_N = \frac{1}{N}\sum_{j=1}^N \widetilde{m}_t^j F(\widetilde{\mathbf{x}}_t^j).$$

Since $\{(\widetilde{\mathbf{x}}_t^j, \widetilde{m}_t^j)\}_{j=1}^N$ are i.i.d. samples of μ, we have

$$\mathbb{E}[|\overline{X}_N - \mathbb{E}[X]|^2|\Omega] \leqslant \mathbb{E}[|\overline{X}_N - \mathbb{E}[X]|^2]/\mathbb{P}(\Omega) \leqslant \text{Var}[\overline{X}_N]/\mathbb{P}(\Omega)$$

$$\leqslant \frac{1}{N}\text{Var}[X]/\mathbb{P}(\Omega) = \mathcal{O}\left(\frac{1}{N}\right),$$

where we use $\mathbb{P}(\Omega) \geqslant 1 - \frac{C}{N}$. A similar bound also holds for the second term in (vi). Combining with (5.13), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}G(t) \leqslant C(t)G(t) + \mathcal{O}\left(\frac{1}{N}\right) \quad \text{with} \quad G(0) = 0 \,.$$

Finally, calling Grönwall's inequality, we have $G(t) \leqslant \mathcal{O}\left(\frac{1}{N}\right)$, which completes the proof. □

5.2. Large-time behavior. This section is dedicated to the proof of Theorem 2.4. This theorem quantifies the large-time behavior of the PDE (2.4a). We note that (2.7) and (2.10) are asymptotic and rate specific, respectively. Both proofs are by contradiction. To start off, we first present two lemmas.

LEMMA 5.3. *Under the condition of Theorem* 2.4, *assume* $\liminf_{t\to\infty} \overline{\mathscr{F}}_t^\mu > F_*$, *then there exists $\epsilon_0 > 0$ such that for any $t \geqslant 0$, we have $\overline{\mathscr{F}}_t^\mu > F_* + \epsilon_0$.*

At the same time, we have the following.

LEMMA 5.4. *Under the condition of Theorem* 2.4, *if there is an $\epsilon > 0$ so that $\overline{\mathscr{F}}_t^\mu > F_* + \epsilon$ for all $t \geqslant 0$, then the total mass $\int \int m \mathrm{d}\mu_t(\mathbf{x}, m) \to \infty$ as $t \to \infty$ and is not conserved.*

The proofs of these two lemmas are found in Appendix B.[6] Putting these two lemmas together, we can prove the theorem.

Proof of Theorem 2.4. We prove (2.7) using proof by contradiction. This is to assume (2.7) is not true, meaning $\liminf_{t\to\infty} \overline{\mathscr{F}}_t^\mu > F_*$. Then according to Lemma 5.3, there exists $\epsilon_0 > 0$ such that $\overline{\mathscr{F}}_t^\mu > F_* + \epsilon_0$ for all $t \geqslant 0$. Setting $\epsilon = \epsilon_0$ in Lemma 5.4, mass is not conserved, contradicting the property of the equation. Hence (2.7) has to be true, namely,

$$\liminf_{t\to\infty} \overline{\mathscr{F}}_t^\mu = F_* \,.$$

To show (2.10), we again deploy a proof by contradiction and the argument is rather similar to that of (2.7). Assuming that it is not true, then for all $t \leqslant t_\epsilon(m_c)$,

$$\overline{\mathscr{F}}_t^\mu \geqslant F_* + \epsilon \,.$$

We will show this leads to a contradiction. Define η_ϵ to be the time so that

$$\eta_\epsilon = \arg \sup_{t \in [0,1]} \int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_t(\mathbf{x}, m) \quad \Rightarrow \quad \int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_{\eta_\epsilon}(\mathbf{x}, m) > 0 \,.$$

Since the probability is nontrivial, one can sample a point in the set of $\mathcal{N}_\epsilon \times (m_c, \infty)$. Starting with $\mathbf{x}(\eta_\epsilon) \in \mathcal{N}_\epsilon$ and $m(\eta_\epsilon) > m_c$, with the same calculation as in (B.6), we get

$$\mathbf{x}(t) \in \mathcal{N}_\epsilon, \quad m(t) > e^{\epsilon(t-\eta_\epsilon)/2}m(\eta_\epsilon) \geqslant e^{\epsilon(t-1)/2}m(\eta_\epsilon)$$

for any $t > 1 \geqslant \eta_\epsilon$, which makes

$$\int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_t(\mathbf{x}, m) > e^{\epsilon(t-1)/2} \int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_{\eta_\epsilon}(\mathbf{x}, m) \,.$$

---

[6]The two lemmas require the assumption of a unique global minimum as stated in Assumption 2.3. However, in the proof in Appendix B, we will discuss that this assumption can be relaxed, but analysis needs to be altered.

In particular, we choose $t = t_\epsilon(m_c)$, then according to definition (2.9),

$$\int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_{t_\epsilon(m_c)}(\mathbf{x}, m) > e^{\epsilon(t_\epsilon(m_c)-1)/2} \int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_{\eta_\epsilon}(\mathbf{x}, m) \geqslant 1 \,.$$

This contradicts the mass conservation:

$$\int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m \mathrm{d}\mu_t(\mathbf{x}, m) \leqslant 1 \quad \forall t \geqslant 1 \,.$$

Hence there is a $t \leqslant t_\epsilon(m_c)$ so that $\overline{\mathscr{F}}_t^\mu < F_* + \epsilon$. □

6. Numerical experiments. This section is dedicated to numerical results. We will showcase numerical evidence that validates theoretical statements on deriving the mean-field limit and the long time behavior of $\overline{F}^N$. Furthermore, we will also present some numerical findings that are yet to be understood theoretically.

All numerical experiments were run on two classical nonconvex functions: the Ackley type functions (in 1, 2, and 10 dimensions) and the Rastrigin type functions in 1 dimension, defined as follows:

$$F_A(\mathbf{x}) = -A \exp\left(-C\sqrt{\frac{1}{d}\sum_{i=1}^d x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^d \cos(2\pi x_i)\right) + A + e + D,$$

$$F_R(\mathbf{x}) = 10d + \sum_{i=1}^d [x_i^2 - 10\cos(2\pi x_i)] + D \,.$$

Throughout the numerical section, we chose the following parameters $A = 20$, $C = 0.2$, and $D = 0$. The profiles of the two functions in 1 dimension are presented in Figure 6.1. They are both highly oscillatory, presenting abundant local minima. Moreover, the local minima for the Rastrigin function give very close function values to the global minimum, and can easily confuse samples on convergence.

For the numerical evidence in section 6.1 we will use the following σ function in the application of Algorithm 1.1:

(6.1)
$$\sigma_{\lambda,\beta}(m) = \begin{cases} \lambda \exp\left(\frac{m}{m-\beta}\right), & m < \beta, \\ 0, & m \geqslant \beta. \end{cases}$$
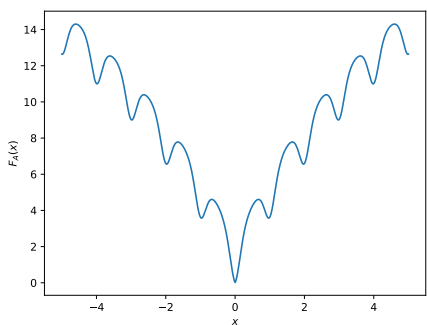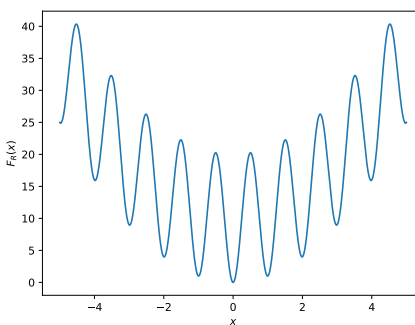
(a) Ackley function.   (b) Rastrigin function.

FIG. 6.1. *Nonconvex functions that will be used for the numerical examples when $d = 1$. Note: color appears only in the online article.*
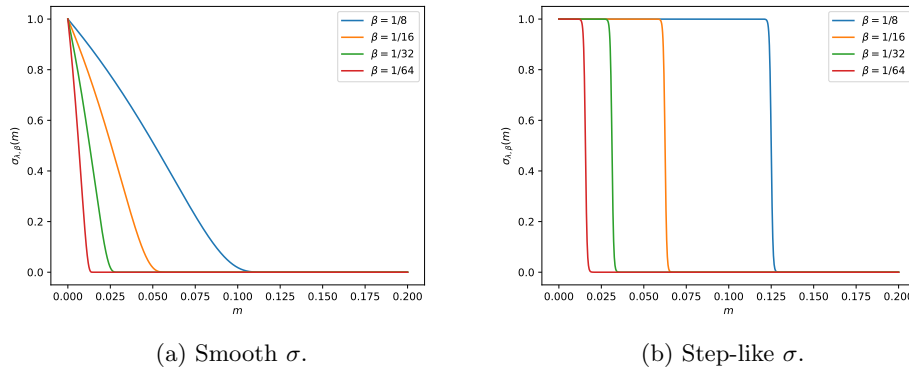
(a) Smooth σ.   (b) Step-like σ.

FIG. 6.2. *σ function for different values of the parameter β and $\lambda = 1$. Note: color appears only in the online article.*

The smooth decay of the function (see Figure 6.2(a)) gives different magnitudes of noise to each particle depending on their mass, but eventually for sufficiently high mass the noise will disappear. We only present three examples in this section. Extensive numerical experiments are presented in Appendix C (see Figures C.12 to C.17).

For the numerical study in section 6.2 we will use the following σ function so we have a stricter change in the application of the noise after a certain amount of mass:

$$(6.2) \qquad \sigma_{\lambda,\beta}(m) = \lambda \left(-\frac{1}{2}\tanh(1000(m - \beta)) + \frac{1}{2} \right).$$

The hyperbolic tangent form of the function is chosen to closely resemble a discontinuous cutoff function for samples whose weight is below β (see Figure 6.2(b)); this means particles with smaller weights are encouraged to explore the landscape of the objective with the same rate λ while all heavier particles are given zero stochasticity. Experimentally we observe a very neat relation between β and N for a consistent convergence rate, and we leave that to section 6.2.

To demonstrate the performance of the algorithm, we choose the experiments where the initial swarm does not cover the global minimum basin, to rule out the easy-to-converge situations. The convergence of the algorithm purely depends on the balancing between Brownian motion and local gradient descent through mass transfer.

6.1. Validation of theoretical results. The first part of our simulation is to showcase the theoretical results we had for Algorithm 1.1. For this, we will apply the algorithm on the Ackley function for $d = 1, 2$, then Rastrigin for $d = 1$, and, finally, to the Ackley function with $d = 10$. In all cases we will show the performance of the algorithm and check its dependence on T and N.

6.1.1. Ackley function in lower dimensions. We experiment with the Ackley function for $d = 1$ and $d = 2$. First, for $d = 1$, we pick $N = 8$ particles, $\beta = 1/8$, $\lambda = 1$, and run the algorithm for step size $h = 10^{-4}$ and 20000 iterations for a total time of $T = 2$. Figure 6.3 demonstrates a few time instances of the particles; It is clear that they gradually move towards the global minimum. We then tested the convergence with $N = 8, 16, 32, 64$. Each run of the algorithm provides one trajectory of $\overline{F}_t^N$ and, in simulation, we ran the algorithm 20 times and compute its expectation, first and third quartiles. The solid lines in Figure 6.4 demonstrate the convergence of $\mathbb{E}[\overline{F}_t^N]$ in time, and the shaded areas show the difference between the first and third quartiles.
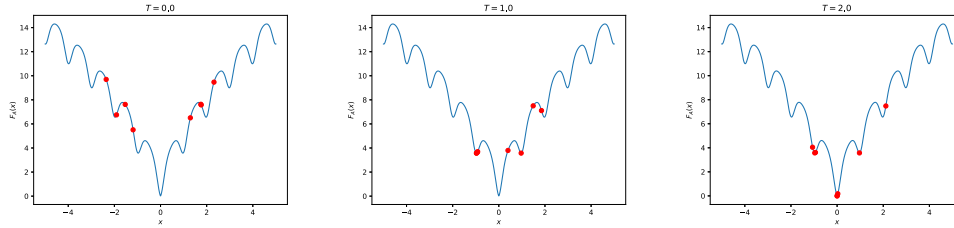
FIG. 6.3. *Swarm movement (in red) for Ackley function in* $d = 1$. *Note: color appears only in the online article.*
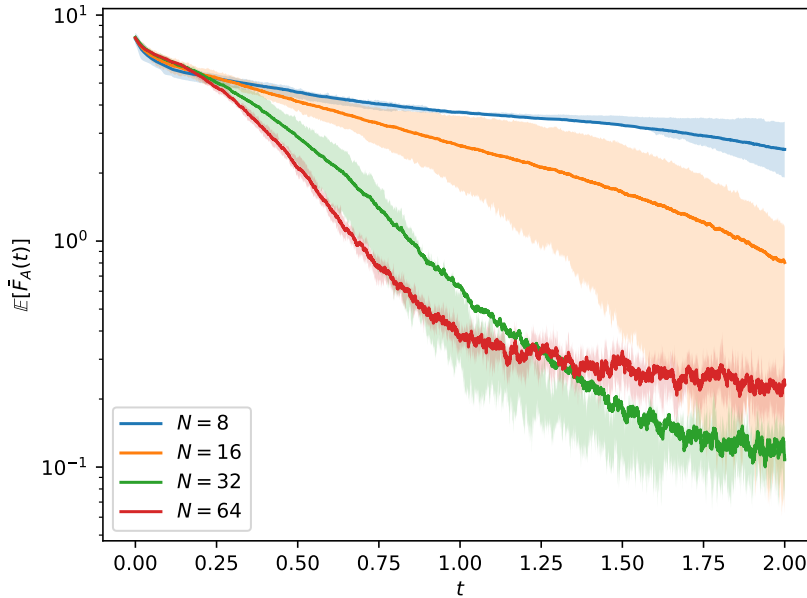

FIG. 6.4. *Expectation of* $\overline{F}_t^N$ *for different values of N for Ackley function in* $d = 1$ *in log-scale. The shaded area shows the difference between the lower quartile and the higher quartile for* $\overline{F}_t^N$. *Note: color appears only in the online article.*
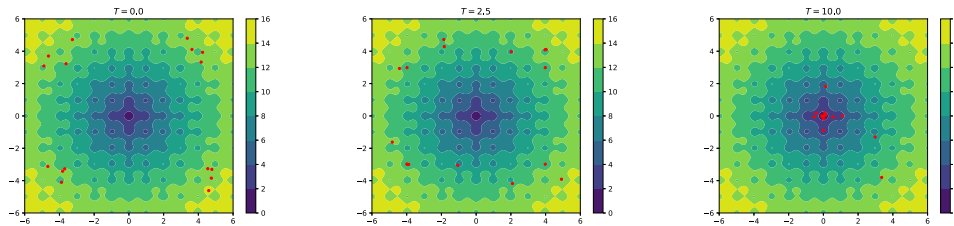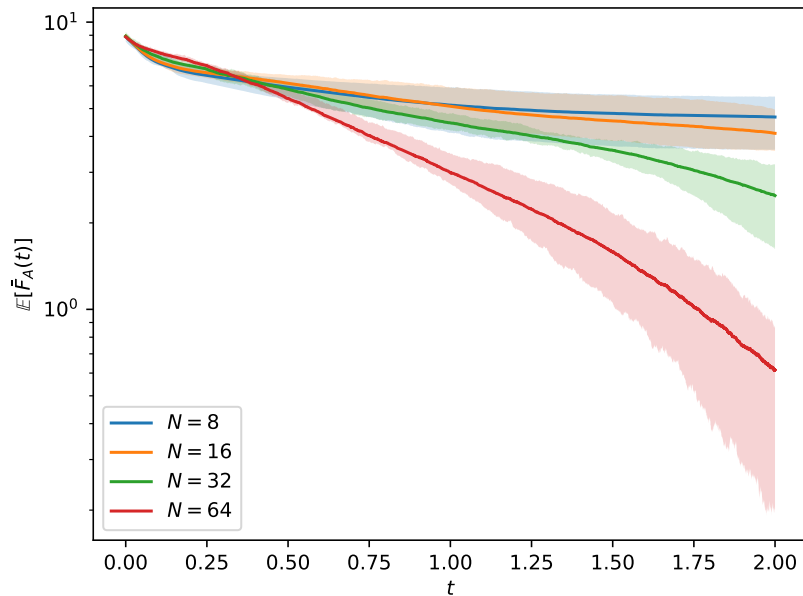

FIG. 6.5. *Swarm movement (in red) for Ackley function in* $d = 2$. *Note: color appears only in the online article.*

For $d = 2$, we pick $N = 16$ particles, $\beta = 1/16$, $\lambda = 1$ and we run the algorithm for step size $h = 10^{-4}$ and 20000 iterations for a total time of $T = 2$. In Figure 6.5 we visualize the convergence of particles to the global minimum in a few time instances. We repeat the experiment for the $d = 1$ case and compute the mean, first and third quartiles of $\overline{F}^N$ for different choices of N, and they are plotted in Figure 6.6.

FIG. 6.6. *Expectation of $\overline{F}_t^N$ for different values of N for Ackley function in $d = 2$ in log-scale. The shaded area shows the difference between the lower quartile and the higher quartile for $\overline{F}_t^N$. Note: color appears only in the online article.*
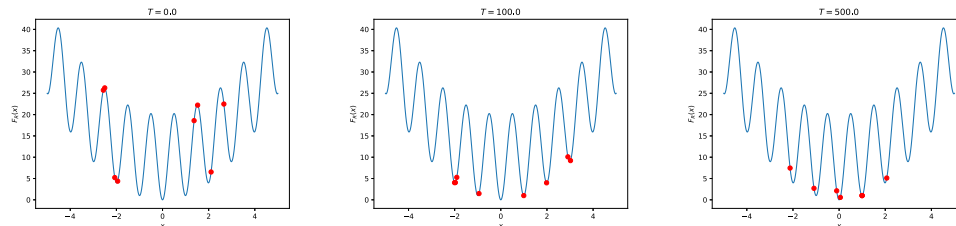

FIG. 6.7. *Swarm movement (in red) for Rastrigin function in $d = 1$. Note: color appears only in the online article.*

6.1.2. Rastrigin function in low dimension. The third test is conducted for the Rastrigin function for $d = 1$. The choice of parameters are $N = 8$, $\beta = 1/8$, and $\lambda = 2$. The step-size is $h = 10^{-4}$ and we run it up to $T = 500$.

In Figure 6.7 we can see that the particles eventually find the global minimum after a long time. The convergence of $\overline{F}_t^N$, however, is significantly faster than finding the global minimum. We also evaluate the mean-field convergence. Running the experiment for 20 runs, we obtain 20 trajectories of $\overline{F}_t^N$. We compute the mean, the first and third quartiles, and plot them in Figure 6.8. It is clear that bigger N gives a higher chance of achieving a lower value for F in time.

6.1.3. Ackley function in higher dimensions. The fourth experiment is to evaluate the performance of the algorithm on a high dimensional function (Ackley function for $d = 10$). The hope is the Brownian motion in the dynamics can help in overcoming the curse of dimensionality. We set the parameters to be $\lambda = 1$, $\beta = 1/N$. In Figure 6.9, we show the convergence in time of 20 runs of experiments using Algorithm 1.1, and compare it to the solution to the deterministic system shown in (4.1). The observation is that the deterministic system converges very quickly at
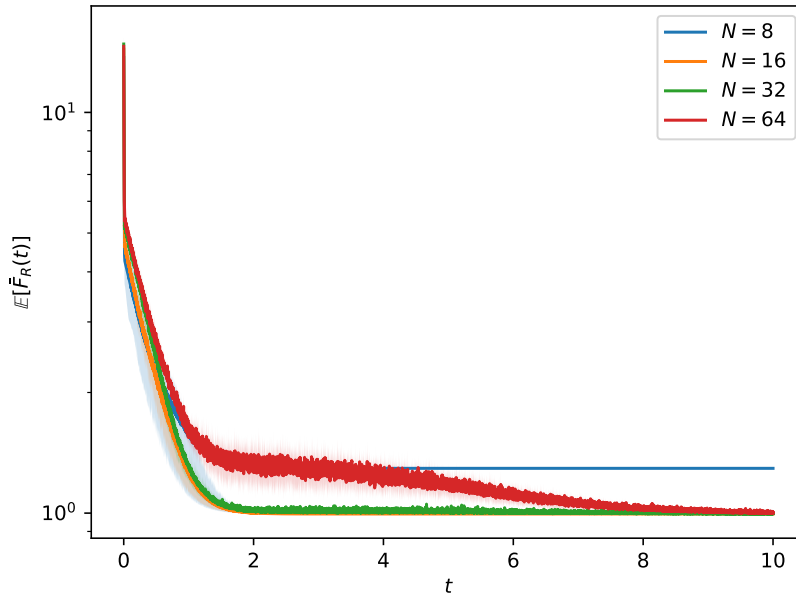
FIG. 6.8. *Expectation of $\overline{F}_t^N$ for different values of N for Rastrigin function in $d = 1$ in log-scale. The shaded area shows the difference between the lower quartile and the higher quartile for $\overline{F}_t^N$. Note: color appears only in the online article.*
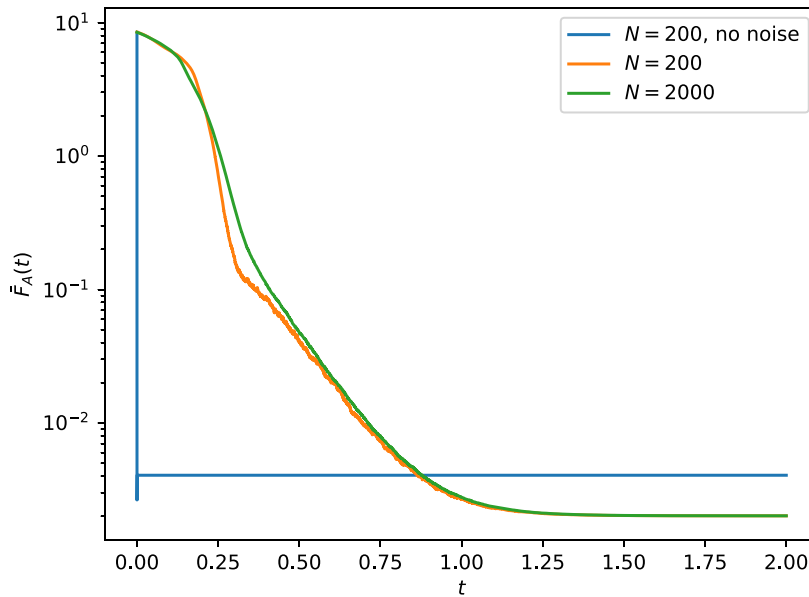

FIG. 6.9. *Long time behavior of $\overline{F}_t^N$ for different values of N for the Ackley function in $d = 10$, in log-scale, in the absence and presence of noise. Note: color appears only in the online article.*

the initial steps, but it saturates at local minima, and the global optimization is never achieved. On the contrary, Algorithm 1.1, using either 2000 or 200 samples, achieves the global minimum for both, albeit with a slower convergence rate. We attribute this to the randomness in the method and this observation confirms Theorem 4.1. We also emphasize that the performance of the algorithm using a smaller number of

samples (200) is compatible to that using 2000, so a small number of samples has already demonstrated impressive numerical performance in this example.

6.2. Numerical study. Though not the focus of the paper, we observe, through running numerical simulations, some interesting algorithmic performances. We collect these findings in this subsection. We first present studies on parameter tuning. As shown earlier, the performance of the algorithm depends on the choice of the σ function that relies on the configuration of the following three parameters: λ, β, and N. While λ determines the strength of the Brownian motion, the cutoff point β and its relation to N fully characterize the percentage of particles that are allowed to "roam." To fully understand the relation, we run Algorithm 1.1 on the Ackley functions for $d = 1$ up to $T = 1$ with σ defined as in (6.2) for $\lambda = 1$ using many choices of (N, β). To be more specific, we choose $N = 8, 16, 32, 64$ and $\beta = 1/8, 1/16, 1/32, 1/64$ and for each (N, β) configuration, we run the experiment 40 times for a collection of 40 trajectories of $\overline{F}_t^N$. The mean and variance of these $\overline{F}_t^N$ for different (N, β) choices are presented, on log-scale in Figure 6.10. We note that experimentwise, it seems that $N = 2/\beta$ seems to give the
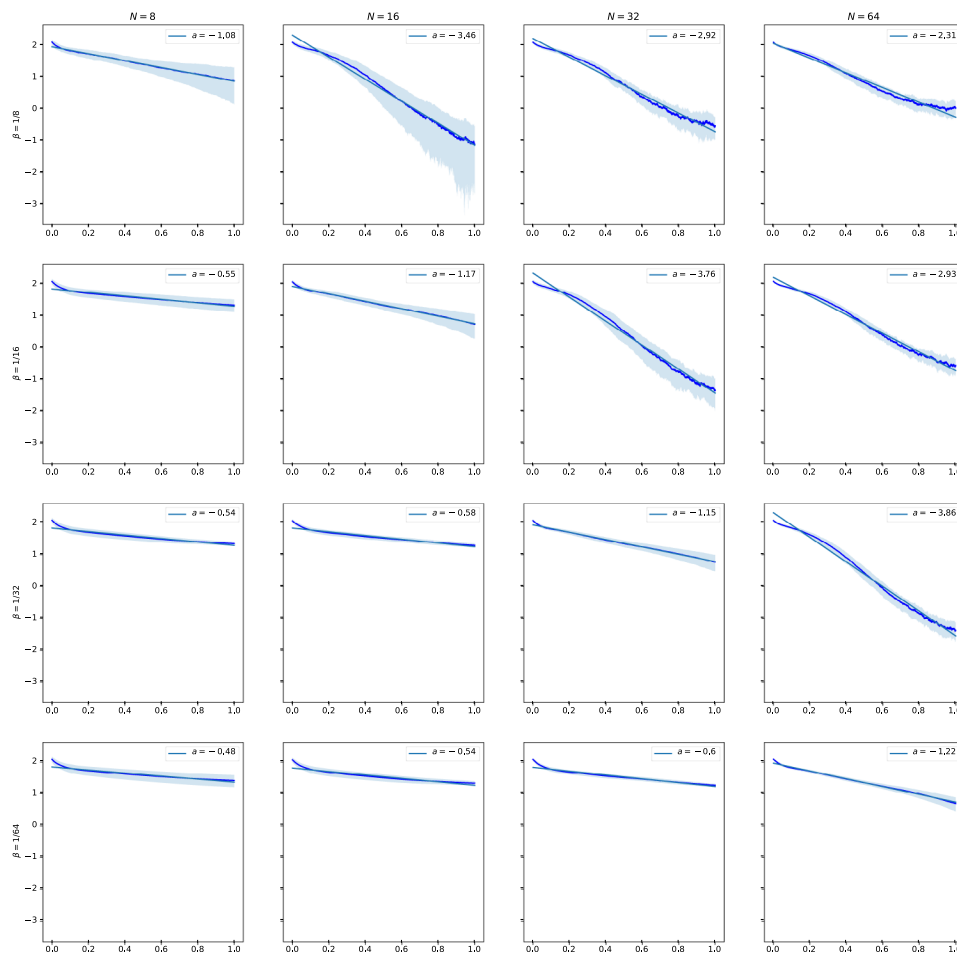

FIG. 6.10. *Convergence rates for the Ackley functions for different values of N and β. Note: color appears only in the online article.*
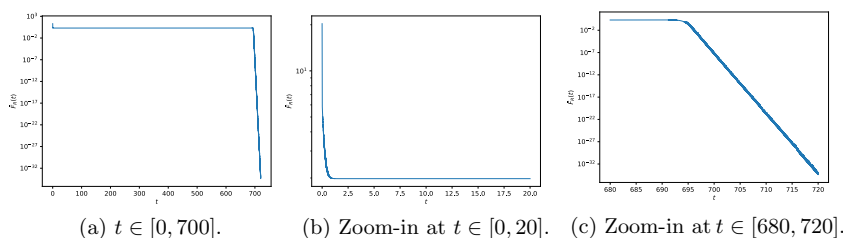
(a) $t \in [0, 700]$.   (b) Zoom-in at $t \in [0, 20]$.   (c) Zoom-in at $t \in [680, 720]$.

FIG. 6.11. *Long time behavior of $\overline{F}_t^N$ for (6.3). Note: color appears only in the online article.*

most ideal convergence. Another observation is that the performance of the algorithm is consistent across the cases when βN are set to be constants. Namely, the convergence rate seem to agree along diagonals in Figure 6.10.

We then present the long time behavior of the algorithm performed on a variation of the Rastrigin function from section 6.1.2,

$$(6.3) \qquad F_R(\mathbf{x}) = 10d + \sum_{i=1}^{d}[2x_i^2 - 10\cos(2\pi x_i)] + D$$

with parameters chosen to be $d = 1$ and $D = 0$, and the same choices of $\{x_i\}$. In this case, we run the simulation for a very long time till $t \sim 700$. As seen in Figure 6.11, the error seems to saturate at around $O(1)$ after $t = 10$, and the second dip occurs at $t = 690$, dragging the error down to 0. The convergence takes place in a staggered pattern. The conjecture is that the value of the local minimum of the Rastrigin function is very close to that of the global one, and the samples stuck at the local minimum for a very long stretch of time, accumulating all the mass. The lighter "explorer" uses this time to explore and find the global minimum. Before the very small basin gets found, the local minimum takes the lead and the error stays at $O(1)$, and the second dip occurs when the global basin finally was found by one explorer, and the mass gets transferred to it. We find this staggered pattern enlightening, and future research is needed for spelling out the explicit dependence of this behavior on different parameters.

Appendix A. Interacting versus noninteracting agents. In this appendix, we investigate the effect of communication, and study the behavior of two systems that produce pitfalls. In subsection A.1 we study basic properties of the deterministic system (4.1) which governs noninteracting agents. In subsection A.2, we switch gears to interacting agents, governed by the stochastic system (4.3) with mass communication passively adjusted according to $\mathbf{x}$-values, and prove Theorem 4.2.

A.1. Properties of the deterministic system (noninteracting agents).

PROPOSITION A.1. *Given the N-particle system defined in (4.1), we have that the provisional minimum $\overline{F}^N(t)$ always decay in time:*[7]

$$(A.1) \qquad \frac{\mathrm{d}}{\mathrm{d}t}\overline{F}^N(t) \leqslant 0 \quad \forall t \geqslant 0, \qquad \overline{F}^N(t) = \frac{1}{M(0)}\sum_{j=1}^{N}m^j(t)F(\mathbf{x}^j(t)).$$

---

[7]Note that the time dependent quantities in the deterministic case are denoted $\square(t, \ldots)$, e.g., $\mathbf{x}^j(t), \overline{F}^N(t), \nu^N(t, \mathbf{x})$, etc.

Proof. This comes from direct computation:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\overline{F}^N(t) &= \frac{1}{M(0)}\left(\sum_{j=1}^{N}\left(\frac{\mathrm{d}}{\mathrm{d}t}m^j(t)\right)F(\mathbf{x}^j(t)) + m^j(t)\frac{\mathrm{d}}{\mathrm{d}t}F(\mathbf{x}^j(t))\right) \\
&= \frac{1}{M(0)}\left(-\sum_{j=1}^{N}m_t^j\left(F(\mathbf{x}^j(t)) - \overline{F}^N(t)\right)F(\mathbf{x}^j(t)) - m^j(t)\left|\nabla F(\mathbf{x}^j(t))\right|^2\right) \\
&= -\frac{1}{M(0)}\sum_{j=1}^{N}m_t^j\left(F(\mathbf{x}^j(t)) - \overline{F}^N(t)\right)^2 - \frac{1}{M(0)}\sum_{j=1}^{N}m_t^j\left|\nabla F(\mathbf{x}^j(t))\right|^2 \leqslant 0.
\end{aligned}
$$

This concludes the proof. □

Similar to its stochastic counterpart, we can also derive the mean-field limiting equation in the deterministic case. This is expressed in terms of the limiting distribution $\nu(t,\mathbf{x})$. To be specific, we have the following proposition.

PROPOSITION A.2. *Let $\{\mathbf{x}^j(t), m^j(t)\}_{j=1}^{N}$ denote the crowd of agents satisfying the deterministic system (4.1) subject to initial data $\{\mathbf{x}^j(0), m^j(0) = \frac{1}{N}\}_{j=1}^{N}$, independently drawn from random distribution ρ_0, and let $\nu^N(t,\mathbf{x})$ denote its empirical distribution*

$$
\nu^N(t,\mathbf{x}) = \sum_{j=1}^{N}m^j(t)\delta_{\mathbf{x}^j(t)}(\mathbf{x}). \tag{A.2}
$$

Then its mean-field limit of the ensemble distribution, $\nu^N(t,\mathbf{x}) \to \nu(t,\mathbf{x})$, satisfies

$$
\partial_t \nu(t,\mathbf{x}) = \nabla_{\mathbf{x}} \cdot (\nu(t,\mathbf{x})\nabla F(\mathbf{x})) - (F(\mathbf{x}) - \overline{\mathscr{F}}^\nu(t))\nu(t,\mathbf{x}), \tag{A.3}
$$

subject to $\nu(0,\cdot) = \rho_0(\cdot)$. Here $\overline{\mathscr{F}}^\nu(t)$, the weighted average, is

$$
\overline{\mathscr{F}}^\nu(t) = \frac{\displaystyle\int F(\mathbf{x})\mathrm{d}\nu(t,\mathbf{x})}{\displaystyle\int \mathrm{d}\nu(t,\mathbf{x})}. \tag{A.4}
$$

Proof. To derive the mean-field limit, we test the system with a test smooth function ϕ:

$$
\mathbb{E}_{\nu(t)}[\phi] = \int \phi(\mathbf{x})\mathrm{d}\nu(t,\mathbf{x}) = \sum_{j=1}^{N}m^j(t)\phi(\mathbf{x}^j(t)).
$$

We take the time derivative on the two sides of the equation. The left-hand side provides

$$
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\nu(t)}[\phi] = \int \partial_t \nu(t,\mathbf{x})\phi(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{A.5}
$$

Here the expectation is with respect to the randomly drawn initial data. The time derivative of the right-hand side yields

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{j=1}^{N}m^j(t)\phi(\mathbf{x}^j(t))\right) \\
&= -\sum_{j=1}^{N}m_t^j\underbrace{\nabla_{\mathbf{x}}\phi(\mathbf{x}^j(t)) \cdot \nabla F(\mathbf{x}^j(t))}_{\dot{\mathbf{x}}=-\nabla F(\mathbf{x}(t))} - \sum_{j=1}^{N}\phi(\mathbf{x}^j(t))\underbrace{m^j(t)(F(\mathbf{x}^j(t)) - \overline{F}(t))}_{\dot{m}=-m(t)(F(\mathbf{x}(t))-\overline{\mathscr{F}}^\nu(t))},
\end{aligned} \tag{A.6}
$$

where we integrated the ODE (4.1). Note that the first term in (A.6) can be written as

$$-\sum_{j=1}^{N} m^j(t)\nabla_{\mathbf{x}}\phi(\mathbf{x}^j(t))\cdot\nabla F(\mathbf{x}^j(t)) = -\mathbb{E}_{\nu(t)}[\nabla_{\mathbf{x}}\phi\cdot\nabla F]$$

$$= \int \phi(\mathbf{x})\nabla_{\mathbf{x}}\cdot(\nu(t,\mathbf{x})\nabla F(\mathbf{x}))\mathrm{d}\mathbf{x}$$

using integration-by-parts, and the second term of (A.6) can be written as

$$\sum_{j=1}^{N}\phi(\mathbf{x}^j(t))m^j(t)(F(\mathbf{x}^j(t))-\overline{\mathscr{F}}^{\nu}(t)) = \mathbb{E}_{\nu(t)}[\phi F]-\mathbb{E}_{\nu(t)}[\phi]\mathbb{E}_{\nu(t)}[F]$$

$$= \int \nu(t,\mathbf{x})\phi(\mathbf{x})F(\mathbf{x})\mathrm{d}\mathbf{x}-\overline{\mathscr{F}}^{\nu}(t)\int \phi(\mathbf{x})\nu(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Plugging the latter into (A.6) and combining it with (A.5) we recover (A.3). □

Some properties of this continuous limit are straightforward. In particular, obvious observations made earlier in our discussion yield the following.

PROPOSITION A.3. *Let ν be the mean-field limit satisfying* (A.3). *There holds that*
- *mass conservation: the total mass, $M(t) := \int \nu(t,\mathbf{x})\mathrm{d}\mathbf{x}$, is conserved in time, $\frac{\mathrm{d}}{\mathrm{d}t}M(t) = 0$;*
- *decrease of provisional mean: the weighted mean $\overline{\mathscr{F}}^{\nu}(t)$ decreases in time, $\frac{\mathrm{d}}{\mathrm{d}t}\overline{\mathscr{F}}^{\nu}(t) \leqslant 0$.*

A.2. Proprieties of the stochastic system (interacting agents). We turn to the proof of Theorem 4.2.

Proof. We may assume without loss of generality that $F_* = 0$, and that the total mass is normalized,

$$\iint m\mathrm{d}\mu_0(\mathbf{x},m) = 1,$$

so that $\rho_t(\cdot) = \int m\mathrm{d}\mu_t(\cdot,m)$ can be viewed as a probability density function in $\mathbf{x}$. Multiplying (4.4) by $mF(\mathbf{x})$ and integrating, we obtain that, with a few integration by parts,

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\mathscr{F}}_t^{\mu} = -\iint m|\nabla F|^2\mathrm{d}\mu_t(\mathbf{x},m) - \iint m(F-\overline{\mathscr{F}}_t^{\mu})^2\mathrm{d}\mu_t(\mathbf{x},m) + \iint m\Delta F\mathrm{d}\mu_t(\mathbf{x},m),$$

and rewriting it using ρ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\mathscr{F}}_t^{\mu} = -\mathbb{E}_{\rho}(|\nabla F|^2) - \mathrm{Var}_{\rho}(F) + \mathbb{E}_{\rho}(\Delta F).$$

Because $\nabla_{\mathbf{x}}F$ is L-Lipschitz according to the first condition in Assumption 2.1 and $F_* = 0$, we obtain

$$|\nabla F|^2 \leqslant 2LF(\mathbf{x}).$$

This implies the first term can be lower bounded by the expectation of $F(\mathbf{x})$:

(A.7) $$-\mathbb{E}_{\rho}(|\nabla F|^2) \geqslant -2L\overline{\mathscr{F}}_t^{\mu}.$$

Since $\mathbf{x}_*$ is the unique global minimum, there exists $\epsilon' > 0$ such that

$$F^{-1}([0,\epsilon']) \subset B_R(\mathbf{x}_*).$$

Now, we lower bound the last two terms:

- Because $F \leqslant D$ according to the second condition in Assumption 2.1 and $F_* = 0$, we obtain

$$\mathrm{Var}_\rho(F) = \mathbb{E}_\rho(F^2) - (\overline{\mathscr{F}}_t^\mu)^2 \leqslant D\overline{\mathscr{F}}_t^\mu \left(1 - \frac{\overline{\mathscr{F}}_t^\mu}{D}\right) \leqslant D\overline{\mathscr{F}}_t^\mu,$$

  where we use Assumption 2.1, $0 \leqslant F \leqslant D$ in the inequalities. This implies

  (A.8) $$-\mathrm{Var}_\rho(F) \geqslant -D\overline{\mathscr{F}}_t^\mu.$$

- Because $F(x)$ is ξ-strongly convex in the ball of $B_R(\mathbf{x}_*)$, we obtain that

$$\Delta F \geqslant \xi d \quad \forall \mathbf{x} \in B_R(\mathbf{x}_*)$$

  and because ∇F is L-Lipschitz, we have

$$|\Delta F| \leqslant dL, \forall \mathbf{x} \in \mathbb{R}^d.$$

To lower bound the last term, we first deploy the Markov inequality:

$$\mathbb{P}(\mathbf{x} \in B_R(\mathbf{x}_*)) \geqslant \mathbb{P}(\mathbf{x} \in F^{-1}([0,\epsilon'])) = 1 - \mathbb{P}(F(\mathbf{x}) \geqslant \epsilon') \geqslant 1 - \frac{\overline{\mathscr{F}}_t^\mu}{\epsilon'}.$$

Then, we have

(A.9) $$\mathbb{E}_\rho(\Delta F) \geqslant d\left(\xi \underbrace{\left(1 - \frac{\overline{\mathscr{F}}_t^\mu}{\epsilon'}\right)}_{\text{contribution from } B_R(\mathbf{x}_*)} - \underbrace{L\left(\frac{\overline{\mathscr{F}}_t^\mu}{\epsilon'}\right)}_{\text{contribution outside } B_R(\mathbf{x}_*)}\right).$$

Combining (A.7), (A.8), and (A.9), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\mathscr{F}}_t^\mu \geqslant d\xi - \overline{\mathscr{F}}_t^\mu \left(2L + D + \frac{d\xi}{\epsilon'} + \frac{dL}{\epsilon'}\right).$$

Setting $\epsilon = d\xi/\left(2L + D + \frac{d\xi}{\epsilon'} + \frac{dL}{\epsilon'}\right)$, we obtain that $\frac{\mathrm{d}}{\mathrm{d}t}\overline{\mathscr{F}}_t^\mu \geqslant 0$ if $\overline{\mathscr{F}}_t^\mu \leqslant \epsilon$. This proves (4.5). $\square$

Appendix B. Technical lemmas. We provide proofs of Lemmas 5.3–5.4, and Lemma B.1 that was deployed herein.

Proof of Lemma 5.3. This is obtained by contradiction. Assuming the statement is not true, meaning there is no such ϵ_0 to make $\overline{\mathscr{F}}_t^\mu > F_* + \epsilon_0$, then $\inf_{t \geqslant 0} \overline{\mathscr{F}}_t^\mu = F_*$. Considering $\liminf_{t \to \infty} \overline{\mathscr{F}}_t^\mu > F_*$, there must exist a finite time, denoted as $t_* \geqslant 0$, such that $\overline{\mathscr{F}}_{t_*}^\mu = F_*$. According to Assumption 2.3, F has a unique global minimum at $\mathbf{x}_*$, so at t_*, $\mu_{t_*}(\mathbf{x}, m)$ must have the form of[8]

(B.1) $$\mu_{t_*}(\mathbf{x}, m) = \delta_{\mathbf{x}_*}(\mathbf{x}) \otimes \eta(m),$$

where η is a probability measure over m-dimensions supported on $\mathbb{R}_+$.

---

[8]We note here that if there are multiple global minima $\{\mathbf{x}_k\}_{k=1}^K$, the form of $\mu_{t_*}(\mathbf{x}, m)$ becomes $\mu_{t_*}(\mathbf{x}, m) = \sum_{k=1}^K \delta_{\mathbf{x}_k}(\mathbf{x}) \otimes \eta_k(m)$, where $\sum_k \eta_k$ has measure 1. The rest of the proof stays the same, with $\mathbf{x}_*$ replaced by the collection of $\mathbf{x}_k$. The proof of this lemma depends on Lemma B.1, which can also be revised to accommodate the multiple global minima case.

We will show below that $\mu_{t_*}(\mathbf{x}, m)$ in the form of (B.1) is an equilibrium. If so, $\mu_t(\mathbf{x}, m) = \mu_{t_*}(\mathbf{x}, m)$ for all $t > t_*$ and, thus, $\overline{\mathscr{F}}_t^\mu = F_*$ for all time after t_*. This contradicts the assumption that $\liminf_{t\to\infty} \overline{\mathscr{F}}_t^\mu > F_*$, completing the proof.

To show $\mu_{t_*}(\mathbf{x}, m)$ in the form of (B.1) is an equilibrium amounts to plugging the form into (2.4a) and show the right-hand side vanishes. Since this is a probability-measured solution with a Dirac delta on the $\mathbf{x}$-domain, the proof has to be conducted in the weak sense. Specifically, given any smooth test function $\phi(\mathbf{x}, m)$, we show the right-hand side of (2.4a), when tested by $\phi(\mathbf{x}, m)$, vanishes:

$$
\iint \text{R.H.S}\,(2.4a)\,\big|_{\mu=\delta_{\mathbf{x}_*}(\mathbf{x})\otimes\eta(m)} \times \phi(\mathbf{x}, m)\mathrm{d}\mathbf{x}\mathrm{d}m
$$

(B.2)
$$
= -\underbrace{\iint (F(\mathbf{x}) - \overline{\mathscr{F}}_t^\mu)m\partial_m\phi(\mathbf{x}, m)\mathrm{d}\mu_t(\mathbf{x}, m)}_{\text{Term I}}
$$

$$
- \iint \left(\underbrace{\nabla_{\mathbf{x}}\phi(\mathbf{x}, m)\cdot\nabla F(\mathbf{x})}_{\text{Term II}} + \underbrace{\sigma(m)\Delta_{\mathbf{x}}\phi(\mathbf{x}, m)}_{\text{Term III}} \right)\mathrm{d}\mu_t(\mathbf{x}, m)\,.
$$

We will show all three terms are zero. Indeed,
- since μ_t takes on the form of (B.1),

(B.3)
$$
\text{Term I} = \iint (F(\mathbf{x}) - \overline{\mathscr{F}}_t^\mu)m\partial_m\phi(\mathbf{x}, m)\mathrm{d}\mu_t(\mathbf{x}, m)
$$
$$
= \iint (F_* - F_*)m\partial_m\phi(\mathbf{x}, m)\mathrm{d}\eta(m) = 0\,;
$$

- Similarly,

(B.4)
$$
\text{Term II} = \iint \nabla_{\mathbf{x}}\phi(\mathbf{x}, m)\cdot\nabla F(\mathbf{x})\delta_{x_*}(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\eta(m)
$$
$$
= \nabla F(\mathbf{x}_*)\cdot\int \nabla_{\mathbf{x}}\phi(\mathbf{x}_*, m)\mathrm{d}\eta(m) = 0\,;
$$

- To show Term III $= 0$, we note that

(B.5)
$$
\text{Term III} = \iint \sigma(m)\Delta_{\mathbf{x}}\phi(\mathbf{x}, m)\delta_{x_*}(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\eta(m)
$$
$$
= \int \sigma(m)\Delta_x\phi(\mathbf{x}_*, m)\mathrm{d}\eta(m) = 0\,,
$$

where in the last equation, we used Lemma B.1 and that $\text{Supp}(\eta(m)) \subset [m_c, \infty)$.

Plugging (B.3)–(B.5) into (B.2), it is clear that $\mu_{t_*}(\mathbf{x}, m) = \delta_{\mathbf{x}_*}(\mathbf{x}) \otimes \eta(m)$ is an equilibrium, and thus $\mu_t(\mathbf{x}, m) = \mu_{t_*}(\mathbf{x}, m)$ for all $t > t_*$, making $\overline{\mathscr{F}}_t^\mu = F_*$ and contradicting the assumption that $\liminf_{t\to\infty} \overline{\mathscr{F}}_t^\mu > F_*$. $\qquad\square$

Proof of Lemma 5.4. According to Assumption 2.3, the initial distribution satisfies $(m_c, \infty) \cap \text{supp}_m(\mu_0(\mathbf{x}, m)) \neq \emptyset$, so using the diffusivity property in $\mathbf{x}$-direction of the Fokker–Planck equation, for $\mathcal{N}_{\epsilon/2}$ as defined in (2.8), there must exist $T_c > 0$ such that

$$
\int_{\mathcal{N}_{\epsilon/2}} \int_{m_c}^{\infty} \mathrm{d}\mu_{T_c}(\mathbf{x}, m) > 0\,,
$$

which means there is no trivial probability at T_c to find a sample in the set of $\mathcal{N}_{\epsilon/2} \times (m_c, \infty)$. Denote one such sample $(\mathbf{x}_s(T_c), m_s(T_c))$, we are to show that

(a) $(\mathbf{x}_s, m_s)(t) \in \mathcal{N}_{\epsilon/2} \times (m_c, \infty)$ for all $t > T_c$;

(b) $m_s(t) > \exp(\frac{\epsilon}{2}(t - T_c))m_s(T_c) > \exp(\frac{\epsilon}{2}(t - T_c))m_c$ for any $t > T_c$.

Indeed, we recall its trajectory governed by the SDE:

$$\begin{cases} \mathrm{d}\mathbf{x}_s(t) &= -\nabla F(\mathbf{x}_s(t))\mathrm{d}t + \sqrt{2\sigma(m_s(t))}\mathrm{d}W_t, \\ \frac{\mathrm{d}}{\mathrm{d}t}m_s(t) &= -m_s(t)\left(F(\mathbf{x}_s(t)) - \overline{\mathscr{F}}_t^\mu\right). \end{cases}$$

Since $\mathbf{x}_s(T_c) \in \mathcal{N}_{\epsilon/2}$ and $m_s(T_c) > m_c$, then noting $\sigma(m_s(T_c)) = 0$, we obtain

$$\left.\frac{\mathrm{d}\mathbf{x}_s(t)}{\mathrm{d}t}\right|_{t=T_c} = -\nabla F(\mathbf{x}_s(t)), \quad \left.\frac{\mathrm{d}m_s(t)}{\mathrm{d}t}\right|_{t=T_c} > \frac{\epsilon}{2}m_s(t),$$

where we used the fact that $\overline{\mathscr{F}}_t^\mu > F_* + \epsilon$ while $F(\mathbf{x}_s(T_c)) < F_* + \epsilon/2$ to obtain the equation for m_s. This dynamics suggests that $F(\mathbf{x}_s(t))$ keeps decreasing so $\mathbf{x}_s \in \mathcal{N}_{\epsilon/2}$, and $m_s(t)$ keeps increasing with an exponential rate:

$$m_s(t) > \exp\left(\frac{\epsilon}{2}(t - T_c)\right)m_s(T_c) > m_c,$$

confirming (a)–(b). Accordingly,

(B.6)

$$\int_{\mathcal{N}_{\epsilon/2}} \int_{m_c}^\infty m\mathrm{d}\mu_t(\mathbf{x}, m) = \mathbb{E}(m(t)|\mathbf{x}(t) \in \mathcal{N}_{\epsilon/2}, m(t) > m_c)\mathbb{P}\left(\left\{\mathbf{x}(t) \in \mathcal{N}_{\epsilon/2}, m(t) > m_c\right\}\right)$$

$$\geqslant \mathbb{E}(m(t)|\mathbf{x}(T) \in \mathcal{N}_{\epsilon/2}, m(T) > m_c)\mathbb{P}\left(\left\{\mathbf{x}(T) \in \mathcal{N}_{\epsilon/2}, m(T) > m_c\right\}\right)$$

$$\geqslant \mathbb{E}\left(\exp\left(\frac{\epsilon}{2}(t - T)\right)m(T)|\mathbf{x}(T) \in \mathcal{N}_{\epsilon/2}, m(T) > m_c\right)\mathbb{P}\left(\left\{\mathbf{x}(T) \in \mathcal{N}_{\epsilon/2}, m(T) > m_c\right\}\right)$$

$$= \exp\left(\frac{\epsilon}{2}(t - T)\right)\int_{\mathcal{N}_{\epsilon/2}} \int_{m_c}^\infty m\mathrm{d}\mu_T(\mathbf{x}, m),$$

where we use (a) in the first inequality and (b) in the second inequality. This leads to the fact that

$$\limsup_{t\to\infty} \int_{\mathbb{R}^d} \int_0^\infty m\mathrm{d}\mu_t(\mathbf{x}, m) \geqslant \limsup_{t\to\infty} \int_{\mathcal{N}_\epsilon} \int_{m_c}^\infty m\mathrm{d}\mu_t(\mathbf{x}, m) = \infty,$$

completing the proof.   □

LEMMA B.1. *Under the conditions of Theorem* 2.4, *let* $\mu_t(\mathbf{x}, m)$ *solve* (2.4a). *If at a time* t_* *so that* $\mu_{t_*}(\mathbf{x}, m)$ *takes on the form of* $\delta_{\mathbf{x}_*} \otimes \eta(m)$ *for some probability measure* η, *then* $\mathrm{Supp}(\eta(m)) \subset [m_c, \infty)$.[9]

Proof. This can be seen by deploying a special test function. Set the test function $\phi_*(\mathbf{x}, m) = |\mathbf{x} - \mathbf{x}_*|^2 \xi(\mathbf{x})$, where $\xi(\mathbf{x})$ is a compactly supported, nonnegative,

---

[9]We note that in the statement of the lemma, we assume $\mu_{t_*}(\mathbf{x}, m)$ has a form of one Dirac delta located on $\mathbf{x}_*$. In the situation when there are multiple global minima $\{\mathbf{x}_k\}_{k=1}^K$, the proof can be adjusted to handle the form of $\mu_{t_*}(\mathbf{x}, m) = \sum_{k=1}^K \delta_{\mathbf{x}_k} \otimes \eta_k(m)$, with the conclusion being $\mathrm{Supp}(\eta_k(m)) \subset [m_c, \infty)$ for all k. In the proof, we should employ a list of test functions with each taking the form of $\phi_k(\mathbf{x}, m) = |\mathbf{x} - \mathbf{x}_k|^2 \xi_k(\mathbf{x})$.

smooth function such that $\xi(\mathbf{x}) = 1$ in a neighborhood of $\mathbf{x}_*$. Naturally by definition, $\partial_m \phi_*(\mathbf{x}, m) = 0$. Testing it on the mean-field PDE (2.4a), we obtain that

$$
(B.7) \qquad \underbrace{\frac{\mathrm{d}}{\mathrm{d}t} \iint \phi_*(\mathbf{x}, m) \mathrm{d}\mu_{t_*}(\mathbf{x}, m)}_{\text{Term I}}
$$

$$
= - \iint \left(\underbrace{\nabla \phi_*(\mathbf{x}, m) \cdot \nabla F(\mathbf{x})}_{\text{Term II}} + \sigma(m) \Delta \phi_*(\mathbf{x}, m) \right) \mathrm{d}\mu_{t_*}(\mathbf{x}, m) \,,
$$

where $\left(F(\mathbf{x}) - \overline{\mathscr{F}}_t^{\mu} \right) \partial_m(m\mu) \big|_{\mu_{t_*}(\mathbf{x}, m)} = 0$ and is dropped out upon testing, using the same argument that showed (B.3). We will show that both Term I and Term II are zero.

Indeed, denote $A(t) := \iint \phi_*(\mathbf{x}, m) \mathrm{d}\mu_t(\mathbf{x}, m)$. According to the definition of ϕ_*, $A(t) \geqslant 0$ for all t. Meanwhile, $A(t_*) = \int \phi_*(\mathbf{x}_*, m) \mathrm{d}\eta(m) = 0$. As a consequence,

$$
\text{Term I} = \frac{\mathrm{d}}{\mathrm{d}t} \iint \phi_*(\mathbf{x}, m) \mathrm{d}\mu_{t_*}(\mathbf{x}, m) = \frac{\mathrm{d}}{\mathrm{d}t} A(t_*) = 0 \,.
$$

Similarly Term II $= 0$ is due to the fact that $\nabla F(\mathbf{x}_*) = 0$.

These together suggests $\iint \sigma(m) \Delta \phi_*(\mathbf{x}, m) \mathrm{d}\mu_{t_*}(\mathbf{x}, m) = 0$. Considering $\mu_{t_*}(\mathbf{x}, m) = \delta_{\mathbf{x}_*} \otimes \eta(m)$ and that $\Delta \phi_*(\mathbf{x}_*) = 2d$, we have $\int \sigma(m) \mathrm{d}\eta(m) = 0$, and thus $\mathrm{Supp}(\eta(\cdot)) \subset [m_c, \infty)$.   □

Appendix C. Numerical experiments. In this section we present further numerical evidence of the performance of our SSA Algorithm 1.1 for different types of 2 dimensionsl nonconvex functions. We used (6.1) as our σ function with different parameters and different numbers of particles in each case. These examples are selected from [27], and in each plot below, we show the landscape of the objective function, the convergence of the provisional minimum in time, and four time snapshots of available samples. At the initial time, the samples are purposely removed from the global basin.
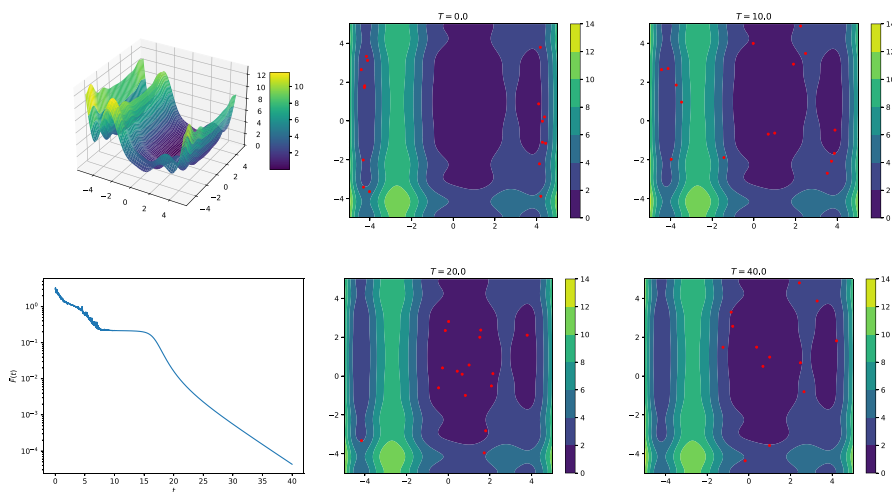

FIG. C.12. *SSA applied to a Levy function, $F(x, y) = \sin^2(\pi w(x)) + (w(x) - 1)^2 [1 + 10 \sin^2(\pi w(x) + 1)] + (w(y) - 1)^2 [1 + \sin^2(2\pi w(y))]$ with $w(x) = 1 + \frac{x-1}{4}$. The parameters used are $N = 16$, $\lambda = 2$, and $\beta = 1/8$. The global minimum is $F(x_*, y_*) = 0$ at $(x_*, y_*) = (1, 1)$. Top left: landscape of he function. Lower left: long time behavior. Center and right: swarm movement in time. Note: color appears only in the online article.*
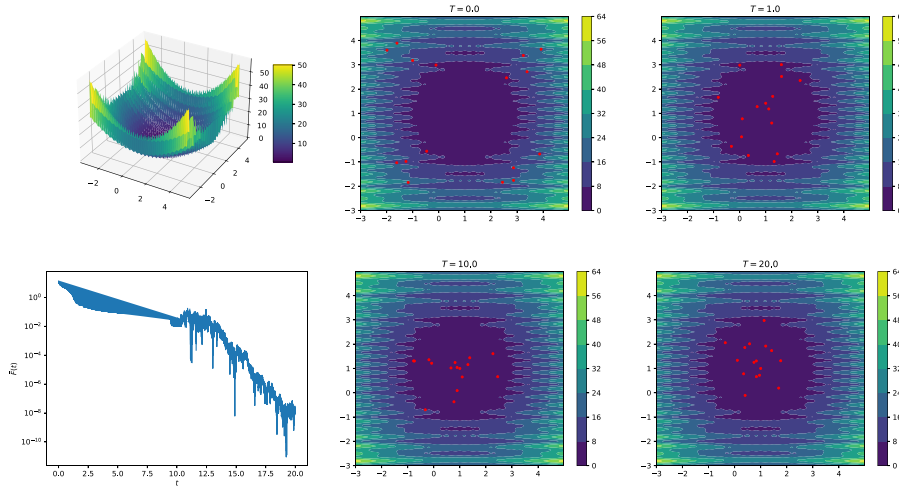
FIG. C.13. *SSA applied to a Levy N.13 function: $F(x,y) = \sin^2(3\pi x) + (x-1)^2[1+\sin^2(3\pi y)] + (y-1)^2[1+\sin^2(2\pi y)]$. The parameters used are $N = 16$, $\lambda = 2$, and $\beta = 1/8$. The global minimum is $F(x_*, y_*) = 0$ at $(x_*, y_*) = (1,1)$. Top left: landscape of the function. Lower left: long time behavior. Center and right: swarm movement in time. Similarly to Figure C.1, we can see the provisional minimum converges to the global minimum as the particles are able to accumulate around the global minimum. Note: color appears only in the online article.*
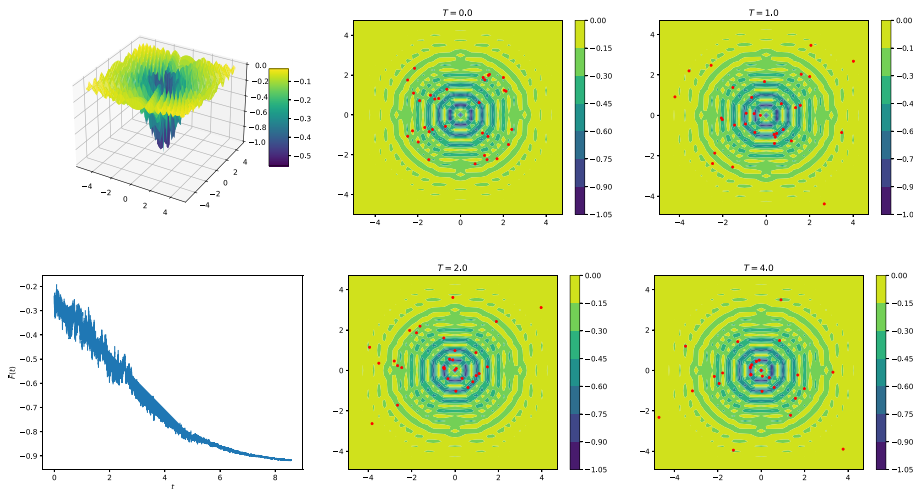
FIG. C.14. *SSA applied to a drop-wave function: $F(x,y) = -\frac{1+\cos(12\sqrt{x^2+y^2})}{0.5(x^2+y^2)+2}$. The parameters used are $N = 32$, $\lambda = 2$, and $\beta = 1/16$. The global minimum is $F(x_*, y_*) = -1$ at $(x_*, y_*) = (0,0)$. Top left: landscape of the function. Lower left: long time behavior. Center and right: swarm movement in time. This is a challenging example since the global basin is very small. Note: color appears only in the online article.*
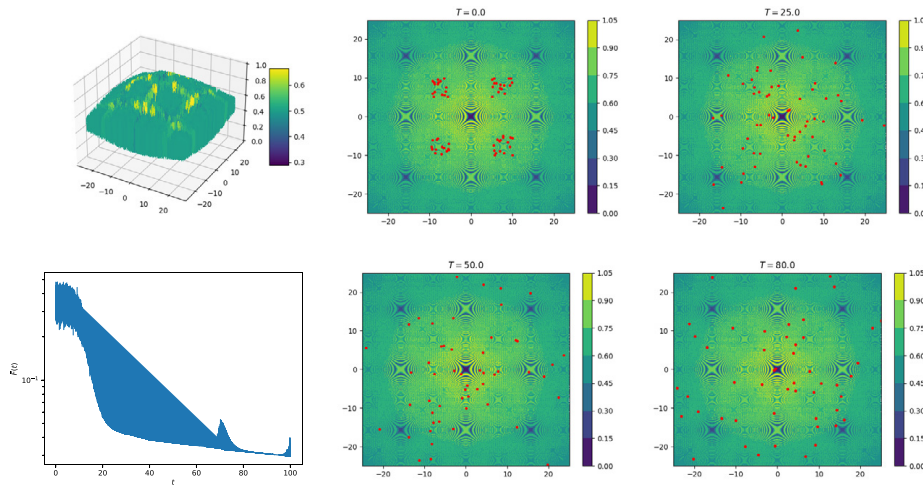
Fig. C.15. *SSA applied to a Schaffer N.2 function:* $F(x,y) = 0.5 + \frac{\sin^2(x^2-y^2)-0.5}{[1+0.001(x^2+y^2)]^2}$. *The parameters used are* $N = 64$, $\lambda = 2$, *and* $\beta = 1/32$. *The global minimum is* $F(x_*,y_*) = 0$ *at* $(x_*,y_*) = (0,0)$. *Top left: landscape of the function. Lower left: long time behavior. Center and right: swarm movement in time. This is also a challenging example since the global basin is very small. Note: color appears only in the online article.*
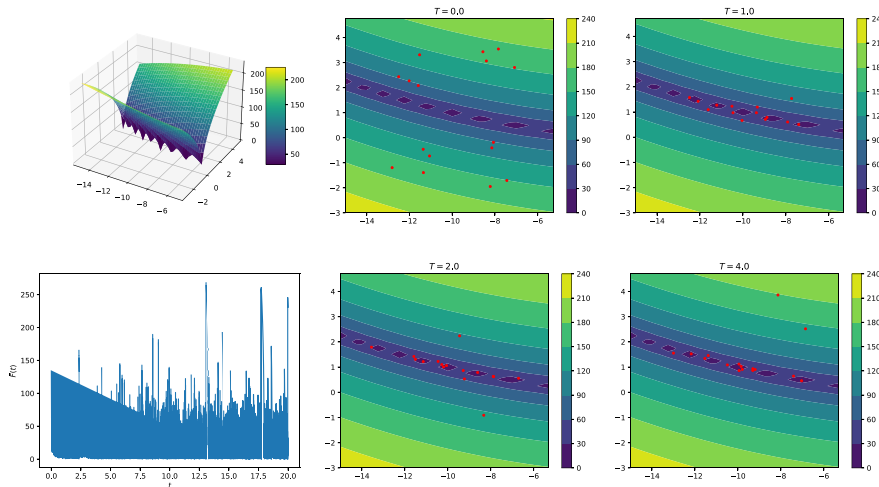

Fig. C.16. *SSA applied to a Bukin N.6 function:* $F(x,y) = 100\sqrt{|y - 0.01x^2|} + 0.01|x + 10|$. *The parameters used are* $N = 16$, $\lambda = 0.25$, *and* $\beta = 1/16$. *The global minimum is* $F(x_*,y_*) = 0$ *at* $(x_*,y_*) = (-10,1)$. *Top left: landscape of the function. Lower left: long time behavior. Center and right: swarm movement in time. This is a very special example: it is observed that the particles can reach the global basin with ease but since all local and global basins are very shallow, samples with Brownian motion easily jump out. As a consequence, the provisional minimum oscillates very wildly. Note: color appears only in the online article.*
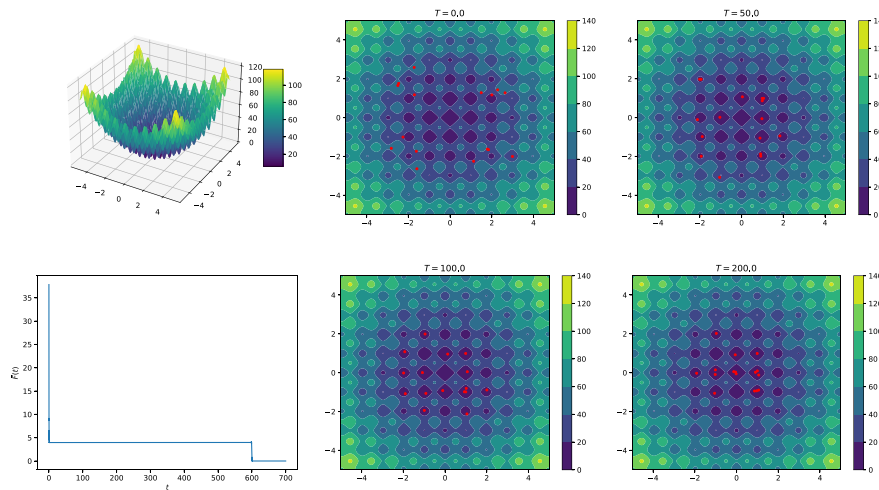
FIG. C.17. *SSA applied to a Rastrigin function: $F(x,y) = 20 + 2x^2 + 2y^2 - \cos(2\pi x) - \cos(2\pi y)$. The parameters used are $N = 16$, $\lambda = 2$, and $\beta = 1/8$. The global minimum is $F(x_*, y_*) = 0$ at $(x_*, y_*) = (0, 0)$. Top left: landscape of the function. Lower left: long time behavior. Center and right: swarm movement in time. Similarly to the 1D Rastrigin case (see Figure 6.11), the provisional minimum quickly decreases to a plateau, and stays there for a long time frame before descending again. The major challenge in this example is that the local minimum has a very similar value to the global minimum, and often, when found, tricks the swarm to saturate. Note: color appears only in the online article.*

Acknowledgment. All four authors thank the two anonymous reviewers for constructive and insightful suggestions.

REFERENCES

[1] Z. BAYRAKTAR, M. KOMURCU, J. A. BOSSARD, AND D. H. WERNER, *The wind driven optimization technique and its application in electromagnetics*, IEEE Trans. Antenn. Propag., 61 (2013), pp. 2745–2757, https://doi.org/10.1109/TAP.2013.2238654.

[2] J. A. CARRILLO, Y.-P. CHOI, C. TOTZECK, AND O. TSE, *An analytical framework for consensus-based global optimization method*, Math. Models Methods Appl. Sci., 28 (2018), pp. 1037–1066, https://doi.org/10.1142/S0218202518500276.

[3] J. A. CARRILLO, S. JIN, L. LI, AND Y. ZHU, *A consensus-based global optimization method for high dimensional machine learning problems*, ESAIM Control Optim. Calc. Var., 27 (2021), S5, https://doi.org/10.1051/cocv/2020046.

[4] J. A. CARRILLO, C. TOTZECK, AND U. VAES, *Consensus-based optimization and ensemble Kalman inversion for global optimization problems with constraints*, in Modeling and Simulation for Collective Dynamics, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. 43, 2023, pp. 195–230, https://doi.org/10.1142/9789811266140_0004.

[5] V. ČERNỲ, *Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm*, J. Optim. Theory Appl., 45 (1985), pp. 41–51, https://doi.org/10.1007/BF00940812.

[6] Y. CHEN, J. CHEN, J. DONG, J. PENG, AND Z. WANG, *Accelerating nonconvex learning via replica exchange Langevin diffusion*, in International Conference on Learning Representations, 2023, https://openreview.net/forum?id=SJfPFjA9Fm.

[7] T.-S. CHIANG, C.-R. HWANG, AND S. J. SHEU, *Diffusion for global optimization in $\mathbb{R}^n$*, SIAM J. Control Optim., 25 (1987), pp. 737–753, https://doi.org/10.1137/0325042.

[8] M. FORNASIER, T. KLOCK, AND K. RIEDL, *Consensus-based optimization methods converge globally*, SIAM J. Optim. 34 (2024), pp. 2973–3004.

[9] N. FOURNIER AND A. GUILLIN, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probab. Theory Related Fields, 162 (2015), pp. 707–738, https://doi.org/10.1007/s00440-014-0583-7.

[10] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6 (1984), pp. 721–741, https://doi.org/10.1109/TPAMI.1984.4767596.

[11] S. GEMAN AND C.-R. HWANG, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031–1043, https://doi.org/10.1137/0324060.

[12] B. GIDAS, *Global optimization via the Langevin equation*, in 1985 24th IEEE Conference on Decision and Control, IEEE, New York, 1985, pp. 774–778.

[13] B. GIDAS, *Nonstationary Markov chains and convergence of the annealing algorithm*, J. Stat. Phys., 39 (1985), pp. 73–131, https://doi.org/10.1007/BF01007975.

[14] S. GRASSI, H. HUANG, L. PARESCHI, J. QIU, P. MARKOVICH, B. PERTHAME, AND E. TADMOR, *Mean-field particle swarm optimization*, in Modeling and Simulation for Collective Dynamics, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. 27, W. Bao, P. Markovich, B. Perthame and E. Tadmor, eds., 2023, pp. 127–193, https://doi.org/10.1142/9789811266140_0003.

[15] D. HENDERSON AND S. H. JACOBSON, *The theory and practice of simulated annealing*, in Handbook of Metaheuristics, Internat. Ser. Oper. Res. Management Sci. 57, Kluwer, Boston, MA, pp. 287–319, 2003.

[16] D. KARABOGA, B. GORKEMLI, C. OZTURK, AND N. KARABOGA, *A comprehensive survey: Artificial bee colony (ABC) algorithm and applications*, Artif. Intell. Rev., 42 (2014), pp. 21–57, https://doi.org/10.1007/s10462-012-9328-0.

[17] J. KENNEDY AND R. EBERHART, *Particle swarm optimization*, in Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4, IEEE, Piscataway, NJ, 1995, pp. 1942–1948.

[18] S. KIRKPATRICK, C. D. GELATT, AND M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680, https://doi.org/10.1126/science.220.4598.671.

[19] P. J. M. LAARHOVEN AND E. H. L. AARTS, *Simulated Annealing: Theory and Applications*, Springer Dordrecht, Netherlands, 1987, pp. 7–15, https://doi.org/10.1007/978-94-015-7744-1.

[20] D. S. LEMONS AND A. GYTHIEL, *Paul Langevin's 1908 paper "on the theory of brownian motion" ["sur la théorie du mouvement brownien," cr acad. sci. (paris) 146, 530–533 (1908)]*, Amer. J. Phys., 65 (1997), pp. 1079–1081, https://doi.org/10.1119/1.18725.

[21] J. LU, E. TADMOR, AND A. ZENGINOGLU, *Swarm-Based Gradient Descent Method for NonConvex Optimization*, arXiv:2211.17157, https://arxiv.org/abs/2211.17157, 2022.

[22] S. MÉLÉARD, *Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models*, in Probabilistic Models for Nonlinear Partial Differential Equations, Springer, Berlin, 1996, pp. 42–95, https://doi.org/10.1007/BFb0093177.

[23] B. C. MOHAN AND R. BASKARAN, *A survey: Ant colony optimization based recent research and implementation on several engineering domain*, Expert Syst. Appl., 39 (2012), pp. 4618–4627, https://doi.org/10.1016/j.eswa.2011.09.076.

[24] P. MONMARCHÉ, *Hypocoercivity in metastable settings and kinetic simulated annealing*, Probab. Theory Related Fields, 172 (2018), pp. 1215–1248, https://doi.org/10.1007/s00440-018-0828-y.

[25] R. PINNAU, C. TOTZECK, O. TSE, AND S. MARTIN, *A consensus-based model for global optimization and its mean-field limit*, Math. Models Methods Appl. Sci., 27 (2017), pp. 183–204, https://doi.org/10.1142/S0218202517400061.

[26] R. POLI, J. KENNEDY, AND T. BLACKWELL, *Particle swarm optimization: An overview*, Swarm Intell., 1 (2007), pp. 33–57, https://doi.org/10.1007/s11721-007-0002-0.

[27] S. SURJANOVIC AND D. BINGHAM, *Optimization test functions and datasets*, in Virtual Library of Simulation Experments: Test Functions and Datasets, https://www.sfu.ca/~ssurjano/optimization.html (2024).

[28] A.-S. SZNITMAN, *Topics in propagation of chaos*, in Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989, P.-L. Hennequin, ed., Springer, Berlin, 1989, pp. 165–251.

[29] E. TADMOR AND A. ZENGINOGLU, *Swarm-based optimization with random descent*, Acta Appl. Math., 190 (2024), https://doi.org/10.1007/s10440-024-00639-0.

[30] C. TOTZECK, *Trends in consensus-based optimization*, in Active Particles, Volume 3: Advances in Theory, Models, and Applications, N. Bellomo, J. A. Carrillo, and E. Tadmor, eds., Springer, Cham, Switzerland, 2021, pp. 201–226.

[31] X.-S. YANG, *Firefly algorithms for multimodal optimization*, in International Symposium on Stochastic Algorithms, Springer, Berlin, 2009, pp. 169–178.