Communications on Applied Mathematics and Computation
https://doi.org/10.1007/s42967-023-00324-3

ORIGINAL PAPER

Adaptive State-Dependent Diffusion for Derivative-Free Optimization

Björn Engquist[1] · **Kui Ren**[2] · **Yunan Yang**[3]

Received: 9 February 2023 / Revised: 14 September 2023 / Accepted: 17 September 2023

Abstract
This paper develops and analyzes a stochastic derivative-free optimization strategy. A key feature is the state-dependent adaptive variance. We prove global convergence in probability with algebraic rate and give the quantitative results in numerical examples. A striking fact is that convergence is achieved without explicit information of the gradient and even without comparing different objective function values as in established methods such as the simplex method and simulated annealing. It can otherwise be compared to annealing with state-dependent temperature.

Keywords Derivative-free optimization · Global optimization · Adaptive diffusion · Stationary distribution · Fokker-Planck theory

Mathematics Subject Classification 90C26 · 90C15 · 65K05

1 Introduction

The idea of using randomness to achieve global convergence in numerical optimization algorithms has been extensively explored. Different stochastic mechanisms have been developed in the literature based on time-dependent diffusion [5, 6, 12, 15, 17, 20]. In [9], we introduced a stochastic gradient descent method for global optimization with a time- and state-dependent variance. Through rigorous analysis of the discrete algorithm and several numerical examples, we demonstrated the global convergence of the algorithm under

✉ Yunan Yang
yunan.yang@cornell.edu

Björn Engquist
engquist@oden.utexas.edu

Kui Ren
kr2002@columbia.edu

[1] Department of Mathematics and the Oden Institute, The University of Texas, Austin, TX 78712, USA

[2] Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA

[3] Department of Mathematics, Cornell University, Ithaca, NY 14853, USA

Published online: 09 January 2024

mild assumptions on the objective function. In this paper, we improve the result in [9] by considering a derivative-free version of the algorithm. We will prove that the new algorithm can still achieve global convergence using a single particle performing Brownian motion where the diffusion coefficient is monotone with respect to the objective function.

To describe the algorithm, let $\Omega \subset \mathbb{R}^d$ $(d \geqslant 1)$ be a smooth bounded domain and $f(\mathbf{x})\colon \Omega \mapsto \mathbb{R}$ a sufficiently regular objective function. We are interested in finding the global minima of f using iterative schemes of the form

$$X_{n+1} = X_n + \sqrt{\eta}\,\sigma\big(f(X_n)\big)\,\boldsymbol{\zeta}_n, \quad n \geqslant 0, \tag{1}$$

where $\{\boldsymbol{\zeta}_n\}_{n\geqslant 0}$ are i.i.d. standard normal random vectors, $\eta > 0$ is the step size, and σ controls the variance of the randomness. Following our previous work [9], we consider adaptive schemes for selecting function-dependent σ values for the iteration. To mimic the classical diffusion setup [17], and also to regularize the degeneracy as done in the literature [29, 34], we introduce a regularization $\epsilon(t) > 0$ with the property that $\epsilon(t) \to 0$ as $t \to \infty$, and define the regularized diffusion coefficient $\sigma = \sigma_\epsilon$ as

$$\sigma_\epsilon(f) = \sqrt{2\Big[\big(f(\mathbf{x}) - f^*_{\min}\big)^+\Big]^\beta + \epsilon(t)}\,, \tag{2}$$

where the exponent $\beta \geqslant d/2$ (d being the dimension of the underlying space), $a^+ := \max\{a, 0\}$, and $f^*_{\min}$ is an approximation to $f_{\min}$, the minimum value of the function $f(\mathbf{x})$ on Ω defined by $f_{\min} := \min_{\mathbf{x}\in\Omega} f(\mathbf{x})$. When $f_{\min}$ is known a priori, we take $f^*_{\min} = f_{\min}$. When $f_{\min}$ is not known, we select $f^*_{\min}$ in other ways (which we will describe in more detail later in the section on numerical simulations). For instance, one choice that has been explored a little bit in [9] and will be investigated more later is the case when $f^*_{\min}$ is taken as the minimum value of f in part of the history of the iteration. That is

$$f^*_{\min} := \min_{n-1-m\leqslant k\leqslant n-1} f(X_k), \quad 1 \leqslant m \leqslant n-1. \tag{3}$$

Without loss of generality, we assume that Ω is a d-dimensional cube with edge length ℓ_Ω, and consider the iteration with periodic boundary condition

$$X_n + \ell_\Omega \mathbf{e}_i = X_n, \quad \forall n \geqslant 0, \quad 1 \leqslant i \leqslant d \tag{4}$$

with $\mathbf{e}_i$ being the unit vector in direction i. We assume that f is periodically extended to $\mathbb{R}^d$ to satisfy $f(\mathbf{x} + \ell_\Omega \mathbf{e}_i) = f(\mathbf{x})$, for any $i, 1 \leqslant i \leqslant d$.

The scheme (1) is a derivative-free stochastic iteration, as it does not explicitly involve the derivative of the objective function f. In the rest of this work, we will show that algorithm (1) with appropriately selected $\epsilon(t)$ on a continuous level and under reasonable assumptions can be globally convergent with an algebraic rate. To be more specific, we show a probability result of the form:

$$\mathbb{P}\big(|X_t - \mathbf{x}_*| > t^{-\nu}\big) \lesssim t^{-\kappa'}, \quad \beta > \frac{d}{2}$$

for some $\nu, \kappa' > 0$; see more details in Theorem 1 and Corollary 1. We will also provide some numerical examples in applications to show its practical relevance; see Sects. 2.3 and 3. Moreover, while our primary focus is to study the derivative-free algorithm (1), we will see that adding explicit gradient information to the algorithm will significantly accelerate its convergence; see Fig. 1.

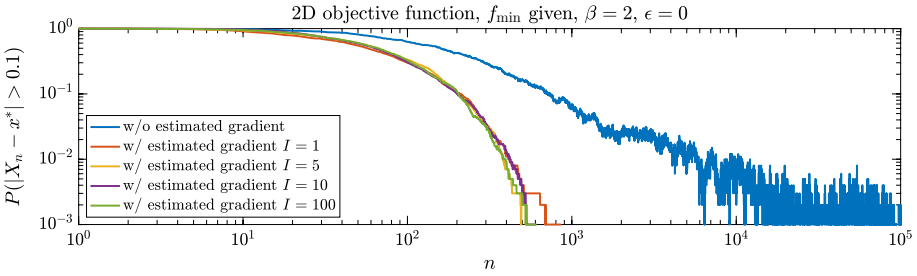


Fig. 1 Log-log plots of convergence performance between (1) and (42). For both cases, we set $D_n(X_n) = f(X_n)^2$ and $f_{\min} = 0$ is known a priori. We also consider different window size I for estimating the gradient in (41). The statistics are estimated by 10^3 i.i.d. runs

There are many effective derivative-free methods in the literature for optimization [3, 7, 23, 25]. While it is impossible to have an exhaustive list of successive methods in this direction, let us mention, as examples, the Nelder-Mead (NM) method [24, 26, 28], which performs a direct search in the parameter space using function value comparison, the genetic algorithm (GA) [18, 21, 27], the simulated annealing (SA) [5, 8, 17, 20, 22], the particle swarm optimization (PSO) method [31, 33], and the consensus-based optimization method [12, 35]. Different variations of such methods have been proposed to solve problems with different features. Interested readers are referred to [1, 7, 23] and references therein for an overview of some of the recent developments in the field. Let us emphasize that some of the methods mentioned above aim at local optimization, and most of them include gradient information implicitly, for instance, by utilizing the difference of objective function values at two different points of the parameter space in the design of the algorithms. However, the scheme (1) does not involve such gradient information in its form. It can be seen as a variant of the Brownian motion where the diffusion coefficient is chosen to depend on the current value of the objective function.

The rest of the paper is structured as follows. In Sect. 2, we present a convergence theory for the algorithm in the continuous limit. We use a regularized version of (2) and assume the value of the global minimum of the objective function is known. In Sect. 2.3, we provide numerical simulations to validate this convergence result. We discuss the algorithm in more practical settings in Sect. 3 for cases where the gradient information can be added, and the objective function value at the global minimum is unknown a priori. We also point out in Sect. 4 the close connection as well as main differences of scheme (1) to our previous work of [9]. Concluding remarks are presented in Sect. 5.

2 Asymptotic Behavior via the Fokker-Planck Equation

We are interested in obtaining a systematic understanding of the algorithm (1). As a starting point, we will analyze this iterative scheme in the continuous limit (whose existence we formally assume, for instance, when $\eta \to 0$ at a proper rate). The iteration is described by the stochastic differential equation (SDE)

$$\mathrm{d}X_t = \sigma(f)\,\mathrm{d}W_t, \tag{5}$$

where $\{W_t\}_{t\geqslant 0}$ is a standard d-dimensional Brownian motion. We formally introduce the generator $\mathcal{L}$ of the process $\{X_t\}_{t\geqslant 0}$ as

$$\mathcal{L}\rho := \frac{1}{2}\sigma^2\Delta\rho, \quad \rho \in \mathcal{C}^2_{\text{per}}(\mathbb{R}^d), \tag{6}$$

where Δ is the standard Laplacian operator in dimension d and the subscript "per" in $\mathcal{C}^2_{\text{per}}(\mathbb{R}^d)$ is used to reflect the fact that functions in the space are Ω-periodic. Then the Fokker-Planck equation for the distribution $u(x, t)$ of the process is of the form

$$\partial_t u = \mathcal{L}^* u := \frac{1}{2}\Delta(\sigma^2 u) \quad \text{in} \quad \mathbb{R}^d \times (0, +\infty), \quad u(\mathbf{x}, 0) = u_0 \quad \text{in} \quad \mathbb{R}^d, \tag{7}$$

assuming that the initial distribution we started the process with, u_0, is also Ω-periodic.

The fact that $\sigma(f^*_{\min}) = 0$ means that the SDE (5), as well as the PDE (7), is *degenerate*. Moreover, as we will see later, our assumption on $f(\mathbf{x})$ and our selection of $\beta \geqslant d/2$ allow singular measures to be admissible solutions to the Fokker-Planck equation (7). These factors make it nontrivial to fully characterize the behavior of the process $\{X_t\}_{t\geqslant 0}$.

We denote by $\mathcal{L}_\varepsilon$ the generator associated with the process with σ_ε in (2). That is,

$$\mathcal{L}_\varepsilon := D_\varepsilon\Delta, \quad D_\varepsilon := \frac{1}{2}\sigma^2_\varepsilon = \left((f(\mathbf{x}) - f^*_{\min})^+\right)^\beta + \varepsilon. \tag{8}$$

We will see later through numerical simulations that the algorithm (1) with adaptive diffusion (2) can be quite efficient in general.

While iteration (1) is derivative-free in nature as it does not explicitly have the gradient of the objective function involved, gradient information is indeed encoded in the algorithm. This can be seen on the heuristic level from the Fokker-Planck equation (7). Indeed, after a little rearrangement, the equation can be written as

$$\partial_t u = \nabla \cdot (u\,\nabla\sigma) + \frac{1}{2}\sigma(f)\,\Delta u - \frac{1}{2}(\Delta\sigma)u. \tag{9}$$

At a given function value f, the first three terms of this Fokker-Planck equation correspond to the SDE (5) with an additional drift term $-\nabla\sigma\,\mathrm{d}t$ on the right-hand side. The drift term $\nabla\sigma = \sigma'(f)\nabla f$ (and $\sigma'(f) > 0$ under the assumptions) clearly depends on the gradient of the objective function. The last term, $-\frac{1}{2}(\Delta\sigma)u$, adds an absorption/generation mechanism in the process at locations where σ is convex/concave.

We first provide some theoretical investigations of our algorithm in the case where the value of the global minimum of f, denoted by $f_{\min}$, is known a priori. In this case, we take $f^*_{\min} = f_{\min}$ in (2). We make the following assumptions on the objective function f.

- A1 The function $f(\mathbf{x})$ is at least $\mathcal{C}^2$ and is Ω-periodic with a unique global minimizer $\mathbf{x}_* \in \Omega$ with $f_{\min} := f(\mathbf{x}_*)$. Moreover, there is a gap $\mathfrak{g}$ between the global minimum value $f_{\min}$ and other local minima of $f(\mathbf{x})$.
- A2 There exist $r, a > 0$ such that $f(\mathbf{x}) - f_{\min} \leqslant a|\mathbf{x} - \mathbf{x}_*|^2$ on $\mathcal{B}_r(\mathbf{x}_*) := \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x} - \mathbf{x}_*| < r\} \subset \Omega$.
- A3 There exists $b > 0$ such that $f(\mathbf{x}) - f_{\min} \geqslant b|\mathbf{x} - \mathbf{x}_*|^2$ for all $\mathbf{x} \in \Omega$.

Remark 1 The rationale for making some of the assumptions in A1–A3 is mainly to simplify the presentation, as it will be evident from the discussions in the rest of this section

that these assumptions can be relaxed significantly for the main results to remain valid. For example, the theoretical result will hold if we replace A1–A3 with the following.

B1 The function $f(\mathbf{x})$ is Ω-periodic with $K < +\infty$ global minimizers $\{\mathbf{x}_k\} \subset \Omega$ with $f_{\min} := f(\mathbf{x}_k)$, for any k, $1 \leqslant k \leqslant K$. Moreover, there is a gap $\mathfrak{g}$ between the global minimum value $f_{\min}$ and other local minima of $f(\mathbf{x})$.

B2 There exist $r, a > 0$ such that for each $1 \leqslant k \leqslant K$, $\left(f(\mathbf{x}) - f_{\min}\right)^{\beta} \leqslant a|\mathbf{x} - \mathbf{x}_k|^{d_*}$ on $\mathcal{B}_r(\mathbf{x}_k) := \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x} - \mathbf{x}_k| < r\} \subset \Omega$ for some $d_* \geqslant d$.

B3 There exists $b > 0$ such that for each $1 \leqslant k \leqslant K$, $\left(f(\mathbf{x}) - f_{\min}\right)^{\beta} \geqslant b|\mathbf{x} - \mathbf{x}_*|^{d_*}$ for all $\mathbf{x} \in \Omega$ and some $d_* \geqslant d$.

In particular, B2 and B3 say that the behavior of $(f - f_{\min})^{\beta}$ is an essential component of the analysis. This makes the theory work for a larger class of objective functions. We will provide numerical simulations in Sect. 2.3 to illustrate the case with multiple global minimizers.

Our main results will be based on the analysis of the Fokker-Planck equation associated with the generator $\mathcal{L}_\varepsilon$ with regularization of the form:

$$\varepsilon(t) = (1 + t)^{-\alpha}, \quad t \geqslant 0 \tag{10}$$

for some $\alpha > 0$. That is,

$$\partial_t u = \Delta\left(D_\varepsilon u\right) \quad \text{in} \quad \mathbb{R}^d \times (0, +\infty), \quad u(\mathbf{x}, 0) = u_0(\mathbf{x}) \quad \text{in} \quad \mathbb{R}^d, \tag{11}$$

where all quantities involved are Ω-periodic, and D_ε is defined in (8). We are interested in solutions representing probability distributions, so we additionally require the normalization condition

$$\int_\Omega u(\mathbf{x}, t)\mathrm{d}\mathbf{x} = 1, \quad \forall t \geqslant 0.$$

The main strategy for constructing a solution to (11) is based on the instantaneous equilibrium distribution of the problem with a fix $\varepsilon(t^*)$ for some t^*. For that purpose, for any given $t > 0$, we denote by $\bar{u}(\mathbf{x}, t)$, an Ω-periodic function, that solves

$$\Delta\left(D_\varepsilon(\mathbf{x}, t)\bar{u}(\mathbf{x}, t)\right) = 0 \tag{12}$$

with the normalization condition $\int_\Omega \bar{u}(\mathbf{x}, t)\mathrm{d}\mathbf{x} = 1$.

With these assumptions, we can prove the following results.

Theorem 1 *Under Assumptions* A1–A3, *let u and $\bar{u}$ be solutions to* (11) *and* (12), *respectively, for $\varepsilon(t)$ given in* (10). *Take $\beta \geqslant \frac{d}{2}$ and $\alpha \in \left(0, \frac{1}{2}\right] \cap \left(0, \frac{2\beta}{d+3\beta}\right)$. Then there exists $t_0 > 0$ such that for all $t > t_0$, we have*

$$\left\|u(x,t)-\bar{u}(x,t)\right\|_{L^2(\mu)} \lesssim t^{-\gamma}, \qquad \gamma = 1-\left(\frac{d}{2\beta}+\frac{3}{2}\right)\alpha > 0, \tag{13}$$

where $\|\cdot\|_{L^2(\mu)}$ *denotes the weighted* L^2 *norm with measure* $\mathrm{d}\mu = D_\varepsilon(\mathbf{x},t)\,\mathrm{d}\mathbf{x}$.

Theorem 1 yields the following corollary which states that the process $\{X_t\}_{t\geqslant 0}$ generated by (5) with σ_ε given in (2) and ε given in (10) converges in probability to the global minimizer $\mathbf{x}_*$ of $f(\mathbf{x})$.

Corollary 1 *Let* $\beta > \frac{d}{2}$. *Then, under the same setting as in Theorem 1, for any* $\delta > 0$, *we have that*

$$\mathbb{P}\left(|X_t-\mathbf{x}_*|>\delta\right) \lesssim t^{-\kappa}, \quad \kappa = \min\left\{\gamma, \left(1-\frac{d}{2\beta}\right)\alpha\right\} \tag{14}$$

for all $t > t_0$. *Moreover, if we take* $\delta = t^{-\nu}$ *with* ν *such that* $0 < \nu < \min\{\gamma, (\frac{1}{2}-\frac{d}{4\beta})\alpha\}$, *then we have*

$$\mathbb{P}\left(|X_t-\mathbf{x}_*|>t^{-\nu}\right) \lesssim t^{-\kappa'}, \quad \kappa' = \min\left\{\gamma-\nu, \left(1-\frac{d}{2\beta}\right)\alpha-2\nu\right\} \tag{15}$$

for all $t > t_0$.

Remark 2 When $\beta = \frac{d}{2}$, we obtain the standard logarithmic convergence as $\mathbb{P}\left(|X_t-\mathbf{x}_*| > \delta\right) \lesssim (\log t)^{-1}$ after applying Theorem 1.

The rest of this section is devoted to the proof of these results.

2.1 Preliminaries in the Case of Fixed ε

The solution $\bar{u}$ of (12), which we refer to as the instantaneous equilibrium distribution, is the equilibrium solution for the problem (11) with a fixed $\varepsilon > 0$. It is straightforward to verify that, when $D_\varepsilon^{-1} \in L^1(\Omega)$, $\bar{u}$ is given as

$$\bar{u}(\mathbf{x}) = Z_{\bar{u}}^{-1} D_\varepsilon^{-1} = Z_{\bar{u}}^{-1} \frac{1}{\left(f(\mathbf{x})-f_{\min}\right)^\beta + \varepsilon}, \qquad Z_{\bar{u}} := \|D_\varepsilon^{-1}\|_{L^1(\Omega)}. \tag{16}$$

Note that periodicity and non-negativity of $\bar{u}(\mathbf{x})$ force out solutions of the form $D_\varepsilon^{-1}(\boldsymbol{A}\cdot\mathbf{x}+B)$ for some vector $\boldsymbol{A}$ and constant B.

We first show that $\bar{u}(\mathbf{x})$ is well defined for any fixed $\varepsilon > 0$. This requires us to show that $Z_{\bar{u}}$ is finite, in which case $\bar{u}(\mathbf{x}) \geqslant 0$ and $\int_\Omega \bar{u}(\mathbf{x})\mathrm{d}\mathbf{x} = 1$. We have the following lemma.

Lemma 1 *Under Assumptions* A1–A3, *for any given* $\varepsilon > 0$, *we have that*

$$0 < Z_{\bar{u}} \leqslant \frac{5}{2} V_\Omega\, \varepsilon^{-1}$$

with V_Ω *the volume of* Ω. *Moreover, when* ε *is sufficiently small, we have that*

$$Z_{\bar{u}} \geqslant \begin{cases} C_1\, \varepsilon^{-\frac{2\beta-d}{2\beta}}, & \beta > \frac{d}{2}, \\ C_2\, \log\left(\frac{1}{\varepsilon}\right), & \beta = \frac{d}{2} \end{cases} \tag{17}$$

for some positive constants C_1 and C_2 independent of ε.

Proof Let r, $\mathfrak{g}$, a, and b be defined as in Assumptions A1–A3, and define $\phi := (f - f_{\min})^{\beta}$. We first derive the upper bound to show that $\bar{u}(\mathbf{x})$ is well-defined. We observe first, using the notation $\phi_{\leqslant\varepsilon} := \{\mathbf{x}\colon \phi(\mathbf{x}) \leqslant \varepsilon\}$ and $\phi^c_{\leqslant\varepsilon} := (\{\mathbf{x}\colon \phi(\mathbf{x}) \leqslant \varepsilon\})^c = \{\mathbf{x}\colon \phi(\mathbf{x}) > \varepsilon\}$, that

$$\begin{aligned} Z_{\bar{u}} &= \int_{\Omega} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} = \int_{\mathcal{B}_r(\mathbf{x}_*)} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} + \int_{\mathcal{B}_r(\mathbf{x}_*)^c} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{B}_r(\mathbf{x}_*)} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} + \int_{\mathcal{B}_r(\mathbf{x}_*)^c\cap\phi_{\leqslant\varepsilon}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} + \int_{\mathcal{B}_r(\mathbf{x}_*)^c\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x}. \end{aligned} \tag{18}$$

The first term is bounded by $\varepsilon^{-1}V_{\mathcal{B}_r(\mathbf{x}_*)}$ with $V_{\mathcal{B}_r(\mathbf{x}_*)}$ the volume of the ball $\mathcal{B}_r(\mathbf{x}_*)$. By the assumption on $f(\mathbf{x})$, the set $\phi_{\leqslant\varepsilon}$ is compact. Therefore, the second term is bounded by $\varepsilon^{-1}V_{\mathcal{B}_r(\mathbf{x}_*)^c\cap\phi_{\leqslant\varepsilon}} \leqslant \varepsilon^{-1}V_{\Omega}$. To bound the last term, we use the assumptions on f to get

$$\int_{\mathcal{B}_r(\mathbf{x}_*)^c\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} \leqslant \int_{\mathcal{B}_r(\mathbf{x}_*)^c\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{2\varepsilon}\mathrm{d}\mathbf{x} \leqslant \frac{V_{\Omega}}{2\varepsilon}.$$

We can now combine the three terms to get the upper bound of $Z_{\bar{u}}$.

To derive the lower bounds in (17), we assume that $\varepsilon < f_{\min} + \mathfrak{g}$. We observe from (18) that

$$\begin{aligned} Z_{\bar{u}} &\geqslant \int_{\mathcal{B}_r(\mathbf{x}_*)} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{B}_r(\mathbf{x}_*)\cap\phi_{\leqslant\varepsilon}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} + \int_{\mathcal{B}_r(\mathbf{x}_*)\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} \geqslant \frac{1}{2\varepsilon}\int_{\mathcal{B}_r(\mathbf{x}_*)(r)\cap\phi_{\leqslant\varepsilon}} \mathrm{d}\mathbf{x} + \frac{1}{2}\int_{\mathcal{B}_r(\mathbf{x}_*)\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{\phi}\mathrm{d}\mathbf{x}. \end{aligned}$$

Let r_ε be such that

$$a^{\beta} r_{\varepsilon}^{2\beta} = \varepsilon, \quad \text{or equivalently,} \quad r_\varepsilon = a^{\frac{-1}{2}}\varepsilon^{\frac{1}{2\beta}}.$$

Then, we have, by denoting $c_d = \frac{\pi^{\frac{1}{2}}}{\Gamma(\frac{d}{2}+1)}$ (the volume of the unit ball in $\mathbb{R}^d$), that

$$\int_{\mathcal{B}_r(\mathbf{x}_*)\cap\phi_{\leqslant\varepsilon}} \mathrm{d}\mathbf{x} = \begin{cases} V_{\mathcal{B}_r(\mathbf{x}_*)} = c_d\, r^d, & r_\varepsilon \geqslant r, \\ V_{\phi_{\leqslant\varepsilon}} = c_d\, r_\varepsilon^d = c_d\, a^{\frac{-d}{2}}\varepsilon^{\frac{d}{2\beta}}, & r_\varepsilon < r. \end{cases}$$

Moreover, when $r_\varepsilon < r$, we have that

$$\begin{aligned} \int_{\mathcal{B}_r(\mathbf{x}_*)\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{\phi}\mathrm{d}\mathbf{x} &= \int_{\mathcal{B}_r(\mathbf{x}_*)\cap\mathcal{B}_{r_\varepsilon}(\mathbf{x}_*)^c} \frac{1}{\phi}\mathrm{d}\mathbf{x} \geqslant \int_{\mathcal{B}_r(\mathbf{x}_*)\cap\mathcal{B}_{r_\varepsilon}(\mathbf{x}_*)^c} \frac{1}{a^{\beta}|\mathbf{x}-\mathbf{x}_*|^{2\beta}}\mathrm{d}\mathbf{x} \\ &= \frac{\mathcal{A}(\mathbb{S}^{d-1})}{a^{\beta}}\int_{r_\varepsilon}^{r} s^{d-1-2\beta}\mathrm{d}s = \frac{\mathcal{A}(\mathbb{S}^{d-1})}{a^{\beta}} \begin{cases} \frac{1}{d-2\beta}(r^{d-2\beta} - r_\varepsilon^{d-2\beta}), & \beta > \frac{d}{2}, \\ \log\frac{r}{r_\varepsilon}, & \beta = \frac{d}{2}. \end{cases} \end{aligned}$$

Here, $\mathcal{A}(\mathbb{S}^{d-1})$ denotes the area of the sphere $\mathbb{S}^{d-1}$. We can now put these bounds together and utilize the fact that $\int_{\mathcal{B}_r(\mathbf{x}_*)\cap\phi^c_{\leqslant\varepsilon}} \frac{1}{\phi}\mathrm{d}\mathbf{x} > 0$ when $r_\varepsilon \geqslant r$ to finish the proof.

The above calculation shows that $Z_{\bar{u}}$, the integral of D_ε^{-1} over Ω, blows up as $\varepsilon \to 0$. This is a key feature needed for the distribution $\bar{u}$ to concentrate on the global minimizer $\mathbf{x}_*$ for sufficiently small ε, as we prove in the next lemma.

Lemma 2 *Under Assumptions* A1–A3, *for any given function value* $\bar{f} > f_{\min}$ *and* $\delta > 0$, *there exists* $\varepsilon_0 > 0$ *such that for any* $\varepsilon \leqslant \varepsilon_0$,

$$\int_{\{\mathbf{x}:\, f(\mathbf{x})\leqslant \bar{f}\}} \bar{u}(\mathbf{x})\,\mathrm{d}\mathbf{x} \geqslant 1-\delta,$$

where $\bar{u}$, *depending on* ε, *is defined in* (16).

Proof With the same notation $\phi(\mathbf{x}) := (f(\mathbf{x}) - f_{\min})^\beta$, we observe that

$$\int_{\{\mathbf{x}:\, f(\mathbf{x})\leqslant \bar{f}\}} \bar{u}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \frac{1}{Z_{\bar{u}}}\int_{\{\mathbf{x}:\, f(\mathbf{x})\leqslant \bar{f}\}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} = 1 - \frac{1}{Z_{\bar{u}}}\int_{\{\mathbf{x}:\, f(\mathbf{x})> \bar{f}\}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x}.$$

Meanwhile, it is straightforward to see that

$$\begin{aligned}\frac{1}{Z_{\bar{u}}}\int_{\{\mathbf{x}:\, f(\mathbf{x})> \bar{f}\}} \frac{1}{\phi+\varepsilon}\mathrm{d}\mathbf{x} &\leqslant \frac{1}{Z_{\bar{u}}}\int_{\{\mathbf{x}:\, f(\mathbf{x})> \bar{f}\}} \frac{1}{\phi(\mathbf{x})}\mathrm{d}\mathbf{x}\\ &\leqslant \frac{1}{Z_{\bar{u}}}\int_{\{\mathbf{x}:\, f(\mathbf{x})> \bar{f}\}} \frac{1}{(\bar{f}-f_{\min})^\beta}\mathrm{d}\mathbf{x} \leqslant \frac{1}{Z_{\bar{u}}}\frac{V_\Omega}{(\bar{f}-f_{\min})^\beta}.\end{aligned}$$

By the result of Lemma 1, we have that

$$\frac{1}{Z_{\bar{u}}}\frac{V_\Omega}{(\bar{f}-f_{\min})^\beta} \leqslant \frac{V_\Omega}{(\bar{f}-f_{\min})^\beta}\begin{cases}\frac{1}{C_1}\varepsilon^{\frac{2\beta-d}{2\beta}}, & \beta > \frac{d}{2},\\ \frac{1}{C_2}\left(\log\frac{1}{\varepsilon}\right)^{-1}, & \beta = \frac{d}{2}\end{cases} \tag{19}$$

with C_1 and C_2 given as in (17). It is then clear that we can select $\varepsilon = \varepsilon_0$ small enough to make this term smaller than δ. The rest then follows from the monotonicity of the bound in (19) with respect to ε.

As we can see, Corollary 1, based on a finely-tuned time-dependent ε, provides a more precise characterization of this result. It says that when we tune $\varepsilon(t)$ at the rate of $t^{-\alpha}$, we get that the rate of concentration is $\lesssim t^{-\gamma}$ for some $\gamma > 0$ depending on α.

The calculation in Lemma 2 also suggests that the distribution $\bar{u}$ converges to the delta-measure at $\mathbf{x}_*$. This is indeed the case, as we see from the following lemma.

Lemma 3 *Under Assumptions* A1–A3, *we have that* $\bar{u}(\mathbf{x}) \to \delta(\mathbf{x}-\mathbf{x}_*)$ *weakly as* $\varepsilon \to 0$.

Proof Let $\psi(\mathbf{x}) \in \mathcal{C}^\infty(\overline{\Omega})$ be a given function. We first assume $\beta > \frac{d}{2}$. Define $\zeta := \frac{2\beta-d}{2\beta}$ and take $\eta \in (0,\zeta)$. We then have, using the same decomposition as in the proof of the previous lemma, that

$$\int_\Omega \psi(\mathbf{x})\bar{u}(\mathbf{x})\mathrm{d}\mathbf{x} - \psi(\mathbf{x}_*) = \frac{1}{Z_{\bar{u}}}\int_{\phi_{\leqslant \varepsilon^\eta}} \frac{\psi(\mathbf{x})-\psi(\mathbf{x}_*)}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x} + \frac{1}{Z_{\bar{u}}}\int_{\phi^c_{\leqslant \varepsilon^\eta}} \frac{\psi(\mathbf{x})-\psi(\mathbf{x}_*)}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}. \tag{20}$$

By Assumption A3, when ε is sufficiently small, $\phi(\mathbf{x}) \leqslant \varepsilon^\eta$ implies that $b|\mathbf{x}-\mathbf{x}_*|^2 \leqslant \varepsilon^{\eta/\beta}$, that is $|\mathbf{x}-\mathbf{x}_*| \leqslant b^{-1/2}\varepsilon^{\eta/2\beta}$. This then implies that $|\psi(\mathbf{x})-\psi(\mathbf{x}_*)| \leqslant C|\mathbf{x}-\mathbf{x}_*| \leqslant \widetilde{C}\varepsilon^{\eta/2\beta}$ for some positive constants C and $\widetilde{C}$. Therefore, the first term on the right-hand side can be bounded as

$$\begin{aligned}\left|\frac{1}{Z_{\bar{u}}}\int_{\phi_{\leqslant \varepsilon^\eta}} \frac{\psi(\mathbf{x})-\psi(\mathbf{x}_*)}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\right| &\leqslant \frac{1}{Z_{\bar{u}}}\int_{\phi_{\leqslant \varepsilon^\eta}} \frac{|\psi(\mathbf{x})-\psi(\mathbf{x}_*)|}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\\ &\leqslant \|\psi(\mathbf{x})-\psi(\mathbf{x}_*)\|_{L^\infty(\phi_{\leqslant \varepsilon^\eta})}\frac{1}{Z_{\bar{u}}}\int_{\phi_{\leqslant \varepsilon^\eta}} \frac{1}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\\ &\leqslant \|\psi(\mathbf{x})-\psi(\mathbf{x}_*)\|_{L^\infty(\phi_{\leqslant \varepsilon^\eta})} \leqslant \widetilde{C}\varepsilon^{\frac{\eta}{2\beta}}.\end{aligned}$$

The second term on the right-hand side can be bounded as follows:

$$\begin{aligned}\left|\frac{1}{Z_{\bar{u}}}\int_{\phi^c_{\leqslant \varepsilon^\eta}} \frac{\psi(\mathbf{x})-\psi(\mathbf{x}_*)}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\right| &\leqslant \frac{2\|\psi\|_{L^\infty(\Omega)}}{Z_{\bar{u}}}\int_{\phi^c_{\leqslant \varepsilon^\eta}} \frac{1}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\\ &\leqslant \frac{2\|\psi\|_{L^\infty(\Omega)}}{Z_{\bar{u}}}\int_{\phi^c_{\leqslant \varepsilon^\eta}} \frac{1}{\varepsilon^\eta}\mathrm{d}\mathbf{x} \leqslant \frac{2\|\psi\|_{L^\infty(\Omega)}}{Z_{\bar{u}}}\frac{V_\Omega}{\varepsilon^\eta} \leqslant \bar{C}\varepsilon^{\frac{2\beta-d}{2\beta}-\eta},\end{aligned}$$

where we have used in lower bound (17) of $Z_{\bar{u}}$ in the last step. Both terms go to 0 when $\varepsilon \to 0$. Therefore, we have shown that $\int_\Omega \psi(\mathbf{x})\bar{u}(\mathbf{x})\mathrm{d}\mathbf{x} - \psi(\mathbf{x}_*) \to 0$ as $\varepsilon \to 0$. This shows that $\bar{u}(\mathbf{x}) \to \delta(\mathbf{x}-\mathbf{x}_*)$ weakly. The case of $\beta = d/2$ can be proved in exactly the same way if we replace the set $\phi_{\leqslant \varepsilon^\eta}$ by the set $\phi_{\leqslant 1/\sqrt{\log(1/\varepsilon)}}$ in the above calculations. Indeed, when ε is sufficiently small, $\phi(\mathbf{x}) \leqslant (\log(1/\varepsilon))^{-\frac{1}{2}}$ implies that $|\mathbf{x}-\mathbf{x}_*| \leqslant b^{-\frac{1}{2}}(\log(1/\varepsilon))^{-\frac{1}{2\beta}}$ which then implies that $|\psi(\mathbf{x})-\psi(\mathbf{x}_*)| \leqslant \widetilde{C}(\log(1/\varepsilon))^{-\frac{1}{2\beta}}$ (with $\widetilde{C}$ the same as that in the case of $\beta > d/2$). Therefore, we can bound the two terms in the decomposition (20) in this case, respectively, as

$$\left|\frac{1}{Z_{\bar{u}}}\int_{\phi_{\leqslant 1/\sqrt{\log(1/\varepsilon)}}} \frac{\psi(\mathbf{x})-\psi(\mathbf{x}_*)}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\right| \leqslant \|\psi(\mathbf{x})-\psi(\mathbf{x}_*)\|_{L^\infty(\phi_{\leqslant 1/\sqrt{\log(1/\varepsilon)}})} \leqslant \widetilde{C}(\log(1/\varepsilon))^{-\frac{1}{2\beta}},$$

and

$$\left|\frac{1}{Z_{\bar{u}}}\int_{\phi^c_{\leqslant 1/\sqrt{\log(1/\varepsilon)}}} \frac{\psi(\mathbf{x})-\psi(\mathbf{x}_*)}{\phi(\mathbf{x})+\varepsilon}\mathrm{d}\mathbf{x}\right| \leqslant \frac{2\|\psi\|_{L^\infty(\Omega)}}{Z_{\bar{u}}}\frac{V_\Omega}{(\log(1/\varepsilon))^{-\frac{1}{2}}} \leqslant \bar{C}(\log(1/\varepsilon))^{-\frac{1}{2}}.$$

Both terms go to 0 when $\varepsilon \to 0$. The proof is now complete.

2.2 Proofs of Theorem 1 and Corollary 1

We now prove Theorem 1. We split the proof into a few steps.

Lemma 4 *Let u and $\bar{u}$ be solutions to (11) and (12), respectively, and $s(t) := \|u - \bar{u}\|_{L^2(\mu)}$. Then there exists $C > 0$ such that*

$$\frac{\mathrm{d}s}{\mathrm{d}t} \leqslant -C\varepsilon^2 s - \sqrt{V_\Omega}\,\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\,Z_{\bar{u}}^{-1}\,\varepsilon^{-\frac{3}{2}}. \tag{21}$$

Proof We define $v = u - \bar{u}$ and thus $s(t) = \|v\|_{L^2(\mu)} := \left(\int_\Omega v^2 D_\varepsilon \mathrm{d}\mathbf{x}\right)^{\frac{1}{2}}$. It is easy to see that v is Ω-periodic and we have, from (11) and (12), that v solves

$$\partial_t v(\mathbf{x},t) = \Delta\Big(D_\varepsilon(\mathbf{x},t)v(\mathbf{x},t)\Big) - \partial_t \bar{u}(\mathbf{x},t). \tag{22}$$

Multiplying both sides by $D_\varepsilon(\mathbf{x},t)v(\mathbf{x},t)$, and integrating over the spatial domain Ω using the periodic boundary condition then leads to the identity

$$\underbrace{\frac{1}{2}\partial_t\left(\int_\Omega D_\varepsilon\,|v|^2\mathrm{d}\mathbf{x}\right)}_{T_1} + \underbrace{\frac{1}{2}\int_\Omega\left(-\frac{\partial D_\varepsilon}{\partial t}\right)|v|^2\mathrm{d}\mathbf{x}}_{T_2} = \underbrace{-\int_\Omega |\nabla\big(D_\varepsilon v\big)|^2\mathrm{d}\mathbf{x}}_{T_3} \underbrace{-\int_\Omega D_\varepsilon v\,\bar{u}_t\mathrm{d}\mathbf{x}}_{T_4}. \tag{23}$$

We first observe that term T_1 is simply

$$T_1 = \frac{1}{2}\partial_t\Big(\|v\|^2_{L^2(\mu)}\Big) = s(t)\,\frac{\mathrm{d}s}{\mathrm{d}t}. \tag{24}$$

For the term T_2, we observe from (8) and (10) that $\dfrac{\partial D_\varepsilon}{\partial t} = \dfrac{\mathrm{d}\varepsilon(t)}{\mathrm{d}t} = -\alpha\,(1+t)^{-\alpha-1} \leqslant 0$. Therefore, $T_2 \geqslant 0$.

To estimate the term T_3, we will apply the weighted Poincaré inequality in Theorem A1. In our case, we consider the weight

$$w(x) = D_\varepsilon(\mathbf{x},t)^{-1} = \frac{1}{(f(\mathbf{x}) - f_{\min})^\beta + \varepsilon(t)}$$

in the A_2 class (see Definition A1). We can find an upper bound of its A_2 constant $[w]_2$ as

$$\begin{aligned}[w]_2 &= \sup_{Q\subset\mathbb{R}^d}\left(\frac{1}{V_Q}\int_Q w(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)\left(\frac{1}{V_Q}\int_Q w(\mathbf{x})^{-1}\,\mathrm{d}\mathbf{x}\right)\\ &\leqslant \frac{\max\limits_{\mathbf{x}\in\overline{\Omega}} w(\mathbf{x})}{\min\limits_{\mathbf{x}\in\overline{\Omega}} w(\mathbf{x})} = \frac{(f_{\max} - f_{\min})^\beta + \varepsilon(t)}{\varepsilon(t)} \leqslant \frac{(f_{\max} - f_{\min})^\beta + \varepsilon(0)}{\varepsilon(t)} = C_\beta\,\varepsilon^{-1},\end{aligned} \tag{25}$$

where $f_{\max} = \max\limits_{\mathbf{x}\in\overline{\Omega}} f(\mathbf{x})$, and the constant

$$C_\beta = (f_{\max} - f_{\min})^\beta + \varepsilon(0) = (f_{\max} - f_{\min})^\beta + 1 = \max_{\mathbf{x}\in\Omega,\, t\in[0,\infty)} D_\varepsilon(\mathbf{x},t) \tag{26}$$

is independent of time t. Based on (A2), for the hypercube Ω and a Lipschitz function v satisfying $\int_\Omega v\, D_\varepsilon^{-1}\mathrm{d}\mathbf{x} = 0$, we have

$$\int_\Omega |v|^2 D_\varepsilon^{-1}\mathrm{d}\mathbf{x} \leqslant \frac{1}{C\varepsilon}\int_\Omega |\nabla v|^2 D_\varepsilon^{-1}\mathrm{d}\mathbf{x},$$

where the constant $C = \left(C_d^2\, \ell_\Omega^2\, C_\beta\right)^{-1}$ with ℓ_Ω, the edge length of the hypercube Ω, and C_d, the constant introduced in (A2). We therefore have

$$\begin{aligned} T_3 \leqslant & -\int_\Omega |\nabla\big(D_\varepsilon v\big)|^2\, \frac{\varepsilon}{D_\varepsilon}\mathrm{d}\mathbf{x} = -\varepsilon\int_\Omega |\nabla\big(D_\varepsilon v\big)|^2\, D_\varepsilon^{-1}\,\mathrm{d}\mathbf{x} \\ \leqslant & -C\varepsilon^2\int_\Omega |D_\varepsilon v|^2\, D_\varepsilon^{-1}\,\mathrm{d}\mathbf{x} = -C\varepsilon^2 s(t)^2. \end{aligned}$$

The last term T_4 can be bounded from below as follows. We first rewrite the term using $\bar{u} = Z_{\bar{u}} D_\varepsilon^{-1}$ from (16):

$$\begin{aligned} T_4 &= \int_\Omega D_\varepsilon v\, \partial_t\big(Z_{\bar{u}}^{-1} D_\varepsilon^{-1}\big)\mathrm{d}\mathbf{x} \\ &= \int_\Omega v\, \partial_t\big(Z_{\bar{u}}^{-1}\big)\mathrm{d}\mathbf{x} - \int_\Omega v Z_{\bar{u}}^{-1}\,\frac{\mathrm{d}D_\varepsilon}{\mathrm{d}t} D_\varepsilon^{-1}\mathrm{d}\mathbf{x} \\ &= \partial_t\big(Z_{\bar{u}}^{-1}\big)\int_\Omega v\,\mathrm{d}\mathbf{x} - Z_{\bar{u}}^{-1}\,\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\int_\Omega D_\varepsilon^{\frac{1}{2}} v\, D_\varepsilon^{-\frac{3}{2}}\mathrm{d}\mathbf{x} \\ &= -Z_{\bar{u}}^{-1}\,\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\int_\Omega D_\varepsilon^{\frac{1}{2}} v\, D_\varepsilon^{-\frac{3}{2}}\mathrm{d}\mathbf{x}, \end{aligned}$$

where we have used the facts that $\dfrac{\mathrm{d}D_\varepsilon}{\mathrm{d}t} = \dfrac{\mathrm{d}\varepsilon}{\mathrm{d}t}$ and $\displaystyle\int_\Omega v\,\mathrm{d}\mathbf{x} = 0$ for any $t > 0$.

Using the decomposition $v = v^+ - v^-$, where $v^+ = \max\{v, 0\}$ and $v^- = -\min\{v, 0\}$, we have that $|D_\varepsilon^{\frac{1}{2}} v| = D_\varepsilon^{\frac{1}{2}} v^+ + D_\varepsilon^{\frac{1}{2}} v^-$. Moreover, it is easy to check that

$$\begin{aligned} \int_\Omega D_\varepsilon^{\frac{1}{2}} v\, D_\varepsilon^{-\frac{3}{2}}\mathrm{d}\mathbf{x} &\geqslant D_{\max}^{-\frac{3}{2}}\int_\Omega D_\varepsilon^{\frac{1}{2}} v^+\,\mathrm{d}\mathbf{x} - D_{\min}^{-\frac{3}{2}}\int_\Omega D_\varepsilon^{\frac{1}{2}} v^-\,\mathrm{d}\mathbf{x} \\ &\geqslant -D_{\min}^{-\frac{3}{2}}\|D_\varepsilon^{\frac{1}{2}} v\|_{L^1(\Omega)} \geqslant -\frac{\sqrt{V_\Omega}}{\varepsilon^{\frac{3}{2}}}\|v\|_{L^2(\mu)}, \end{aligned}$$

where $D_{\max} := \max\limits_{\mathbf{x}\in\overline{\Omega}} D_\varepsilon$, $D_{\min} := \min\limits_{\mathbf{x}\in\overline{\Omega}} D_\varepsilon$, and $\mathrm{d}\mu = D_\varepsilon\,\mathrm{d}\mathbf{x}$. This leads to

$$T_4 \geqslant Z_{\bar{u}}^{-1}\,\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\,\varepsilon^{-\frac{3}{2}}\,\sqrt{V_\Omega}\,s(t). \tag{27}$$

Finally, we combine all four terms in (23) to have $T_1 = T_3 - T_4 - T_2 \leqslant T_3 - T_4$. The inequality (21) then follows.

We are now ready to prove the main result Theorem 1.

Proof of Theorem 1 Using the lower bound on $Z_{\bar{u}}$ given in Lemma 1, as well as the assumption that $\varepsilon(t) = (1+t)^{-\alpha}$, we can further relax (21) to obtain, after a change of variable $1 + t \to t$,

$$s_t \leqslant -C\, t^{-2\alpha} s + C_2\, t^{\ell\alpha-1}, \quad t \geqslant 1, \quad \ell = \frac{d}{2\beta} - \frac{1}{2}, \tag{28}$$

where C_2 is a positive constant and $\beta \geqslant d/2$.

Next, we find an upper bound for $s(t)$. First, we discuss the case $\alpha \neq 1/2$. Define $C_\alpha = \frac{C}{1-2\alpha}$, and

$$\begin{aligned}
y(t) &:= \frac{1}{2\alpha-1} \exp\left(-\frac{Ct^{1-2\alpha}}{1-2\alpha}\right) = \frac{1}{2\alpha-1} \exp\left(-C_\alpha t^{1-2\alpha}\right),\\
h(t) &:= \Gamma\left(\frac{\ell\alpha}{1-2\alpha}, -C_\alpha\, t^{1-2\alpha}\right) = \int_{-C_\alpha\, t^{1-2\alpha}}^{\infty} \tau^{\frac{\ell\alpha}{1-2\alpha}-1} \mathrm{e}^{-\tau} \mathrm{d}\tau,\\
C_5 &:= C_2\Big(\frac{C}{2\alpha-1}\Big)^{\frac{\ell\alpha}{2\alpha-1}} = C_2\big(-C_\alpha\big)^{\frac{\ell\alpha}{2\alpha-1}},\\
C_3 &:= C_2\Big(\frac{C}{2\alpha-1}\Big)^{\frac{\ell\alpha}{2\alpha-1}} \Gamma\left(\frac{\ell\alpha}{1-2\alpha}, \frac{C}{2\alpha-1}\right) = C_5\, \Gamma\left(\frac{\ell\alpha}{1-2\alpha}, -C_\alpha\right),\\
C_4 &:= (2\alpha-1)s(1) \exp\left(\frac{C}{1-2\alpha}\right) = (2\alpha-1)s(1) \exp\big(C_\alpha\big),
\end{aligned}$$

where $\Gamma(s,x) = \int_x^\infty t^{s-1}\mathrm{e}^{-t}\mathrm{d}t$ is the upper incomplete gamma function. Note that $\Gamma(s,0) = \Gamma(s)$. The solution to the ODE $\bar{s}_t = -Ct^{-2\alpha}\bar{s} + C_2 t^{\ell\alpha-1}$ with the initial condition $s(1) = \bar{s}(1)$ is

$$\bar{s}(t) = (C_4 - C_3)y(t) + C_5\, y(t)\, h(t) \geqslant s(t). \tag{29}$$

If $1-2\alpha < 0$ (i.e., $\alpha > 1/2$), we have

$$y(t) \xrightarrow{t\to\infty} \frac{1}{2\alpha-1}, \qquad h(t) \xrightarrow{t\to\infty} \Gamma\left(\frac{\ell\alpha}{1-2\alpha}\right)$$

both converging to constants. From (29), $\bar{s}(t)$ converges to a constant and we do not have an upper bound decay for $s(t)$ in this analysis framework.

If $1-2\alpha > 0$ (i.e., $0 < \alpha < 1/2$), we have $y(t) \xrightarrow{t\to\infty} 0$ exponentially. Since $-C_\alpha\, t^{1-2\alpha} < 0$ for $t \geqslant 1$, $h(t)$ is a complex-valued scalar with both the real and the imaginary parts going to $-\infty$ as $t \to +\infty$. It is worth noting that $\mathrm{e}^{-x}\Gamma(s,-x) \approx x^{s-1}$ when x is sufficiently large, so when t is large, we have

$$y(t)\, h(t) \approx \frac{1}{2\alpha-1}\big(C_\alpha\, t^{1-2\alpha}\big)^{\frac{\ell\alpha}{1-2\alpha}-1} = C_6\, t^{(\ell+2)\alpha-1},$$

while $y(t) = \frac{1}{2\alpha-1}\mathrm{e}^{-C_\alpha t^{1-2\alpha}} < 0$ based on its definition and C_6 is some positive constant. Thus, $s(t) \leqslant \bar{s}(t) \lesssim t^{(\ell+2)\alpha-1}$. In order for the upper bound to decay to zero, we need $(\ell+2)\alpha - 1 < 0$, i.e.,

$$0 < \alpha < \min\left(\frac{1}{2}, \frac{1}{\ell+2}\right) = \min\left(\frac{1}{2}, \frac{2\beta}{d+3\beta}\right).$$

When $\alpha = 1/2$, we need to consider the ODE

$$\bar{s}_t = -C\,t^{-1}\bar{s} + C_2\,t^{\frac{\ell}{2}-1},$$

where $s(1) = \bar{s}(1)$ as the initial condition. It has an analytical solution. We then have

$$s(t) \leqslant \bar{s}(t) = \frac{2C_2}{2C+\ell}t^{\frac{\ell}{2}} + \frac{s(1)(2C+\ell) - 2C_2}{2C+\ell}t^{-C} \lesssim t^{\frac{\ell}{2}}.$$

If $\ell = \frac{d}{2\beta} - \frac{1}{2} < 0$, i.e., $\beta > d$, we will have an energy decay when $\alpha = 1/2$.

To sum up, when α, β, and d are chosen to satisfy

$$\alpha \in \left(0, \frac{1}{2}\right] \bigcap \left(0, \frac{2\beta}{d+3\beta}\right), \tag{30}$$

we have

$$s(t) \lesssim t^{(\ell+2)\alpha-1} = t^{-\gamma}, \qquad \gamma = 1 - (\ell+2)\alpha.$$

This completes the proof.

The energy estimate in Theorem 1 allows us to refine the result of Lemma 2. This is the result of Corollary 1. We now prove this corollary.

Proof of Corollary 1 For a given $\delta > 0$, we have, based on Assumption A3 and the lower bound estimations for $Z_{\bar{u}}$ in Lemma 1, and after taking into account that $\epsilon(t) = (1+t)^{-\alpha} \sim t^{-\alpha}$ for large t, that

$$\int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} \bar{u}\,\mathrm{d}\mathbf{x} = Z_{\bar{u}}^{-1}\int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} \frac{1}{D_\epsilon(\mathbf{x},t)}\mathrm{d}\mathbf{x} \leqslant Z_{\bar{u}}^{-1}\frac{V_\Omega}{b\delta^2} \lesssim (b\delta^2)^{-1}t^{\alpha(\frac{d}{2\beta}-1)}.$$

On the other hand, we have

$$\begin{aligned}\int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} (u-\bar{u})\mathrm{d}\mathbf{x} &\leqslant \int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} |v|\,\mathrm{d}\mathbf{x} \leqslant \frac{1}{\sqrt{b\delta^2}}\int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} D_\epsilon^{\frac{1}{2}}\,|v|\,\mathrm{d}\mathbf{x}\\ &\leqslant \frac{\sqrt{V_\Omega}}{\sqrt{b\delta^2}}\|D_\epsilon^{\frac{1}{2}}v\|_{L^2(\Omega)} = \frac{\sqrt{V_\Omega}}{\sqrt{b\delta^2}}\,s(t).\end{aligned}$$

Therefore, based on the upper bound estimate for $s(t)$ in Theorem 1, we have

$$\begin{aligned}\mathbb{P}\big(X_t \notin \mathcal{B}_\delta(\mathbf{x}_*)\big) &= \int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} u\,\mathrm{d}\mathbf{x} = \int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} (u-\bar{u})\,\mathrm{d}\mathbf{x} + \int_{\Omega\cap\mathcal{B}_\delta(\mathbf{x}_*)^c} \bar{u}\,\mathrm{d}\mathbf{x}\\ &\leqslant C_1(b\delta^2)^{-\frac{1}{2}}t^{\left(\frac{d}{2\beta}+\frac{3}{2}\right)\alpha-1} + \bar{C}_1(b\delta^2)^{-1}t^{\alpha\left(\frac{d}{2\beta}-1\right)} \lesssim t^{-\kappa},\end{aligned} \tag{31}$$

where $C_1, \bar{C}_1$ are positive constants, and κ is defined in (14). Now if we take $\delta = t^{-\nu}$ with $0 < \nu < \min\{\gamma, (1-\frac{d}{2\beta})\frac{\alpha}{2}\}$, then (31) simplifies to

$$\mathbb{P}\big(X_t \notin \mathcal{B}_\delta(\mathbf{x}_*)\big) \leqslant C_1 b^{-\frac{1}{2}}t^{-(\gamma-\nu)} + \bar{C}_1 b^{-1}t^{-\left[\left(1-\frac{d}{2\beta}\right)\alpha-2\nu\right]} \lesssim t^{-\kappa'},$$

where κ' is defined in (15).

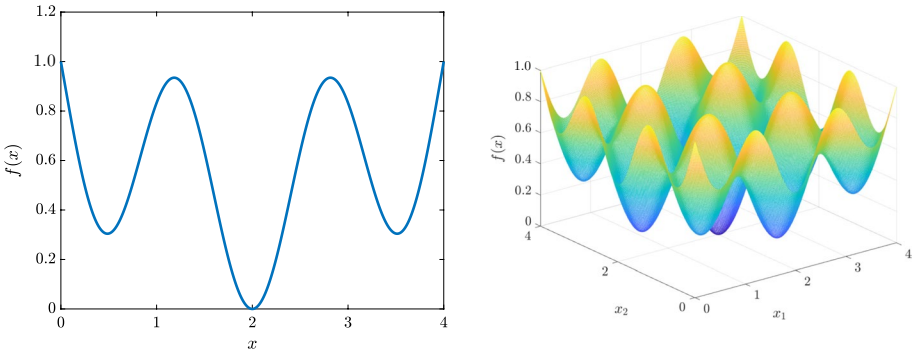


Fig. 2 Optimization landscapes of the objective function $f(\mathbf{x})$ in (32) in dimension $d = 1$ (left) and $d = 2$ (right)

2.3 Numerical Experiments

Next, we show a few numerical examples of global optimization to demonstrate the effectiveness of our proposed derivative-free algorithm.

We will consider minimizing the following objective function $f(\mathbf{x})$ on the domain $\Omega = [0, 4]^d$, where

$$f(\mathbf{x}) = \frac{1}{d\bar{f}}\left(0.3|\mathbf{x} - 2|^2 - \sum_{i=1}^{d} \cos(4x_i - 8) + d\right), \quad \mathbf{x} = [x_1, \cdots, x_d]^\top, \tag{32}$$

where $\bar{f} = 2.345\,5$, so that $f_{\max} = \max_{\mathbf{x}\in[0,4]^d} f(\mathbf{x}) = 1$ for any dimension d. There is a unique global minimum of $f(\mathbf{x})$ at $\mathbf{x}_* = [2, \cdots, 2]^\top \in \Omega$ with the function value $f_{\min} = f(\mathbf{x}_*) = 0$. The shapes of the objective function in dimension $d = 1$ and $d = 2$ are illustrated in Fig. 2.

Our numerical simulations are based on the discrete algorithm (1), where we fix the step size η to be a constant. The standard deviation for the noise is taken as the discrete equivalence of (2), i.e.,

$$\sigma_n = \sqrt{2\left[\left((f(X_n) - f_{\min})^+\right)^\beta + \varepsilon\right]}, \quad \varepsilon = cn^{-\alpha}, \tag{33}$$

where $c = 10^{-3}$ is a fixed scalar. This setup corresponds to the main results of the paper proved in Sect. 2. The update rule for the iterate is

$$X_{n+1} = X_n + \sigma_n \xi_n, \tag{34}$$

where $\xi_n \sim \mathcal{N}(0, I_d)$, the standard normal distribution on $\mathbb{R}^d$ with an enforced periodic boundary condition. The standard deviation σ_n, or equivalently, the diffusion coefficient, is both state- and time-dependent.

The convergence histories are shown in Fig. 3 for the case $d = 2$ with $\beta = 2$, where we vary the value of the parameter α. Based on the log-log plots, we see that the choice of α directly affects the convergence speed in the discrete algorithm as the bigger the α, the faster the convergence. While we will discuss more on the case of $\varepsilon = 0$ in Sect. 3.2, for the purpose of comparison, we include in Fig. 3 a plot for the case where the term $\varepsilon = 0$ as the limit of $\alpha \to \infty$.

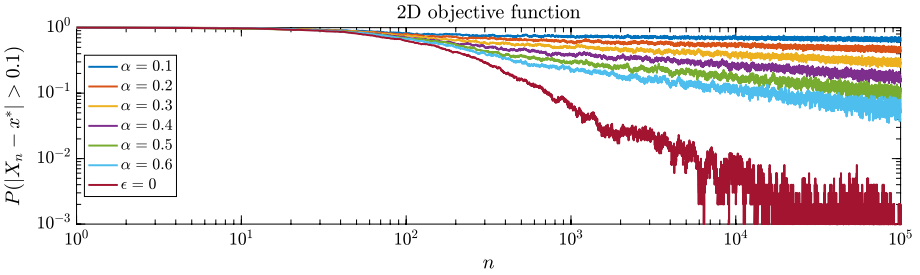


Fig. 3 Convergence history of iteration (1) with σ given in (33) in the case of minimizing $f(\mathbf{x})$ of (32) in dimension $d = 2$ with $\beta = 2$. Shown are results for different values of α after 10^5 iterations

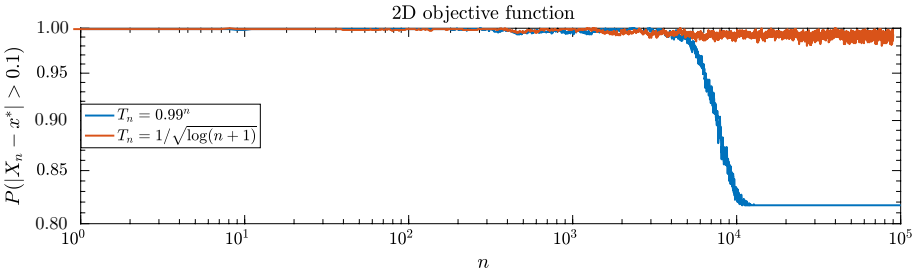
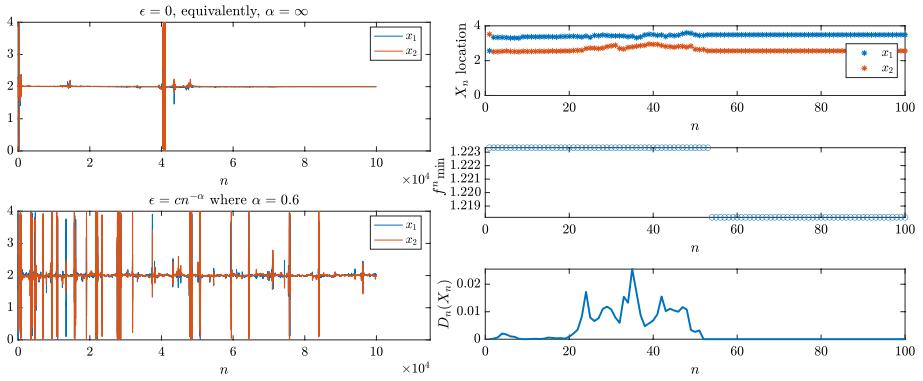


Fig. 4 Convergence history of the SA algorithm [23] to minimize $f(\mathbf{x})$ of (32) in dimension $d = 2$ under two different annealing schedules for the temperature T_n defined in (35)

For comparison, we consider the SA method [22] to minimize the same objective function (32). The algorithm determines whether the next iterate (a close neighbor of the current state) is better or worse than the current iterate and then chooses the next one. If the new iterate yields a lower objective function value, it becomes the next iterate automatically. Otherwise, the SA method accepts the next iterate (a worse point) based on an acceptance probability. Here, we choose the common acceptance probability function

$$\exp\left(-\frac{1}{T_n}\big(f(X_{n+1}) - f(X_n)\big)\right), \tag{35}$$

where the temperature T_n decreases as n increases. The acceptance probability becomes smaller for the fixed function value gap $f(X_{n+1}) - f(X_n)$ as n goes to infinity. This is commonly referred to as the cool-down process. In this comparison, we choose two different types of annealing schedule to decrease T_n: an algebraic decay $T_n = 0.99^n$, and a logarithmic decay $T_n = 1/\sqrt{\log(n+1)}$. In Fig. 4, we plot the same statistical estimates for $\mathbb{P}(|X_n - \mathbf{x}_*| > 0.1)$, similar to Fig. 3. The plots show an interesting phenomenon: when the temperature T_n decays algebraically, the iterates are stuck at local minima after 10^4 iterations; when the temperature decays as slowly as $\mathcal{O}(1/\sqrt{(\log n)})$, there shows no sign of convergence within 10^5 iterations. This is a long-standing dilemma for stochastic global optimization algorithms: too-fast cooling results in local minimum trapping, while too-slow cooling results in slow convergence. The state-dependent diffusion proposed in this



(a) History of the iterates from a single run (b) Trajectory of X_n, $f^n_{\min}$ and $D_n(X_n) = (\sigma_n)^2/2$

Fig. 5 (**a**) The history of the iterates $\{X_n\}$ from a single run when $\varepsilon = 0$ (top) and $\varepsilon = cn^{-\alpha}$ (bottom), respectively, with $f_{\min}$ given. (**b**) The trajectory of the iterates X_n, the estimated minimum value $f^n_{\min}$ and the effective diffusion coefficient $D_n(X_n) = (\sigma_n)^2/2$ from one single run with $\varepsilon = 0$ and $f_{\min}$ unknown. The global minimum is $[2, 2]^\top$ in both cases

work aims to speed-up global convergence by incorporating the objective function into the temperature decay.

3 Practical Generalizations

The numerical results we presented in the previous section verified our theoretical analysis in Sect. 2, where we assumed that the value of the global minimum of the objective function, $f_{\min}$, is known and demonstrated that the algorithm could perform well in more complex situations. In this section, we provide further discussions on practical situations under which our algorithm performs almost as well as in the ideal case.

We start with a numerical illustration for various cases regarding $f_{\min}$ and the value of ε, which will be further discussed in Sects. 3.1 and 3.2, respectively. The left plots of Fig. 5 show single-run trajectories of the cases $\varepsilon = 0$ (top) and $\varepsilon \neq 0$ (bottom), respectively, under the setting that $f_{\min}$ is known (and $f(\mathbf{x}_*) = 0$). The case of $\varepsilon = 0$ is superior in stabilizing the iterates around the global minimum. The right plots of Fig. 5 are results on the trajectory of one single run with $\varepsilon = 0$ when $f_{\min}$ is unknown (and estimated with the method in Sect. 3.1). The iterates get stuck at a wrong position very quickly, showing that we cannot set $\varepsilon = 0$ when $f_{\min}$ is unknown, in contrast to the two left plots. We will elaborate on these observations more in this section.

3.1 Estimating Optimal Objective Function Value

Previously, and also in the main theoretical results, we have used the assumption that the value $f_{\min} := f(\mathbf{x}_*)$ (but not the location $\mathbf{x}_*$) is known a priori. This is often true in many applications (for instance, data matching) where $f_{\min} = 0$. When $f_{\min}$ is unknown, developing a convergence theory for the algorithm is much more challenging. The difficulty is that the above analysis is in the continuum, and the estimation on $f(\mathbf{x}_*)$ is inherently discrete.

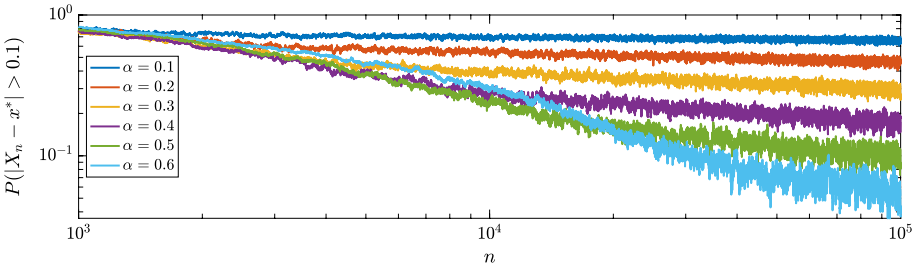


Fig. 6 The convergence performance of (1) with the diffusion coefficient defined in (36) and $\varepsilon = 10^{-3}n^{-\alpha}$. The minimum objective function value $f(\mathbf{x}_*)$ is estimated by $f^n_{\min}$

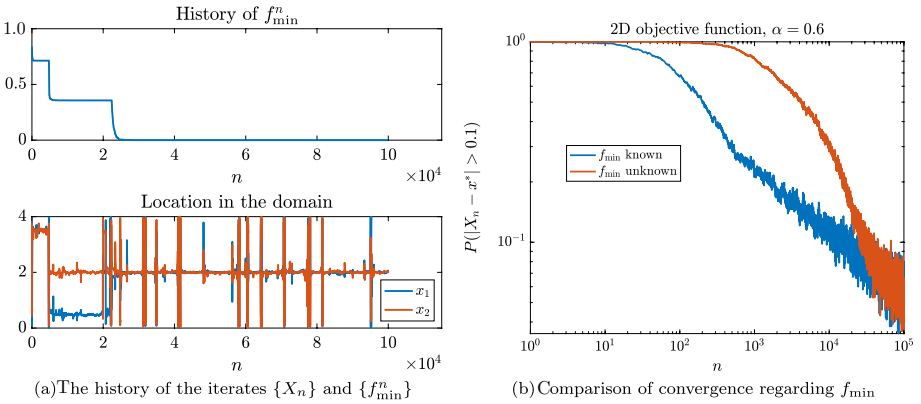


Fig. 7 (**a**) The history of the iterates $\{X_n\}$ and $\{f^n_{\min}\}$ from a single run when $\varepsilon = 10^{-3}n^{-0.6}$ in (36). (**b**) The comparison of convergence performances in terms of whether $f_{\min} = f(\mathbf{x}_*)$ is known or needs to be estimated using $f^n_{\min}$ following (36). In both cases, we set $\varepsilon = 10^{-3}n^{-0.6}$

However, with a little more effort in estimating $f_{\min}$ during the iteration, we can make our algorithm efficient under such a situation.

It is important to have the state-dependent term in the algorithm, which is the only component that encodes any information regarding the objective function $f(\mathbf{x})$. We may consider a different variant of (2) and (33):

$$\sigma_n = \sqrt{2\left[\left(f(X_n) - f^n_{\min}\right)^{\beta} + \varepsilon\right]}, \quad \text{where} \quad f^n_{\min} := \min\left\{f(X_n), f^{n-1}_{\min}\right\}, \ \varepsilon = cn^{-\alpha}. \tag{36}$$

The role of $f^n_{\min}$ here is to approximate $f(\mathbf{x}_*)$ through the history minimum of the objective function values from the past iterates. We also need to have $\varepsilon \neq 0$ not only to avoid $\{X_n\}$ stagnating at any history minimum rather than the global minimum but also to visit everywhere of the domain Ω; see for a counterexample in Fig. 5b.

On the other hand, in the case of $f(\mathbf{x}_*)$ unknown, when we set $\varepsilon \neq 0$ but monotonically decaying as n becomes large, we observe in Fig. 6 the decay of $\mathbb{P}(|X_n - \mathbf{x}_*| > 0.1)$ as n increases, but much slower than the cases shown in Fig. 3 in which we assume to know $f(\mathbf{x}_*) = 0$ *a priori*. In Fig. 7a, we present the history of $f^n_{\min}$ and X_n from a single

run when $\varepsilon = 10^{-3}n^{-0.6}$. In Fig. 7b, we show a comparison regarding whether $f(\mathbf{x}_*)$ is known *a priori* or not where the convergence performances are estimated from 10^3 i.i.d. runs and $\varepsilon = 10^{-3}n^{-0.6}$ in both cases.

The proposed algorithm (36) for estimating the optimal value of the objective function $f_{\min}$ is based on discrete $f(X_n)$ values and does not fit well into the continuum style convergence proof. With increasing values of n, it is possible to approximate $f_{\min} = f(\mathbf{x}_*)$ with increasing accuracy. In the numerical example related to Fig. 7a, the simple estimate (36) of $f(\mathbf{x}_*)$ was used. The figure shows the convergence to $\mathbf{x}_*$ and the optimum estimate to $f(\mathbf{x}_*)$. There are abnormal cases where the estimate (36) would require very slow decay of $\varepsilon(t)$, and for a rigorous convergence result, we adopt the same strategy, which we used in [9], of basing the hyperparameter estimates on extra sampling. If in the sequence in (36) we add uniformly sampled values $\{Y_n\}$ from the domain Ω, we can guarantee almost-sure convergence of $f(Y_n)$ to the optimal value $f(\mathbf{x}_*)$; see Proposition 1.

Proposition 1 *Assume that there is a subset $\Omega_{\text{sc}} \subseteq \Omega$ on which the objective function $f(x)$ is strongly convex and $\mathbf{x}_* \in \Omega_{\text{sc}}$. Define the monotone-decreasing sequence*

$$f^n_{\min} := \min\left\{f(X_n), f(Y_n), f^{n-1}_{\min}\right\}, \quad f^0_{\min} = \min\left\{f(X_0), f(Y_0)\right\},$$

where $\{X_n\}$ are iterates from (34) with $\sigma_n = \sqrt{2(f(X_n) - f^n_{\min})^\beta + 2\varepsilon}$ and $\{Y_n\}$ are uniform samples drawn from the domain Ω. Then we have $f^n_{\min} \xrightarrow{n\to\infty} f(\mathbf{x}_*)$ almost surely.

Proof Let δ be the largest positive constant such that $\Omega_\delta := \{x : f(x) - f(\mathbf{x}_*) \leqslant \delta\} \subseteq \Omega_{\text{sc}}$. Note that Ω_δ is nested between two ellipsoids centered at $\mathbf{x}_*$ and the ratio $|\Omega_\delta|/|\Omega| \leqslant C\delta^{d/2}$ for some positive constant C [9, Eq. (3.22)]. Also, it is easy to see that

$$f^n_{\min} \leqslant \min\left\{f(Y_1), f(Y_2), \cdots, f(Y_n)\right\} := M^{(n)}.$$

Therefore, we have

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty} f^n_{\min} - f(\mathbf{x}_*) > \delta\right) &\leqslant \mathbb{P}\left(\lim_{n\to\infty} M^{(n)} - f(\mathbf{x}_*) > \delta\right) = \mathbb{P}\left(\bigcap_{n=0}^{\infty}\{Y_n \notin \Omega_\delta\}\right) \\ &= \prod_{n=0}^{\infty}\mathbb{P}(Y_n \notin \Omega_\delta) \leqslant \lim_{n\to\infty}\left(1 - C\delta^{d/2}\right)^n = 0.\end{aligned} \tag{37}$$

Since (37) holds for any $0 < \delta' \leqslant \delta$, we conclude that $f^n_{\min}$ converges to $f(\mathbf{x}_*)$ almost surely.

3.2 Regularization-Free Algorithm

In this section, we discuss the case when $\varepsilon = 0$ in (2). Based on the definition in (2), $\sigma(f)$ is not integrable for $\beta \geqslant d/2$. This means that, mathematically, the process $\{X_t\}_{t\geqslant 0}$ can get arbitrarily close to the global minimizer but will never reach it unless $X_0 = \mathbf{x}_*$. However, from a practical point of view, being arbitrarily close is sufficient.

Numerically, we still observe a rapid convergence of the discrete algorithm (1) to the global minimizer in this case, even though this "convergence" might not be in a strict mathematical

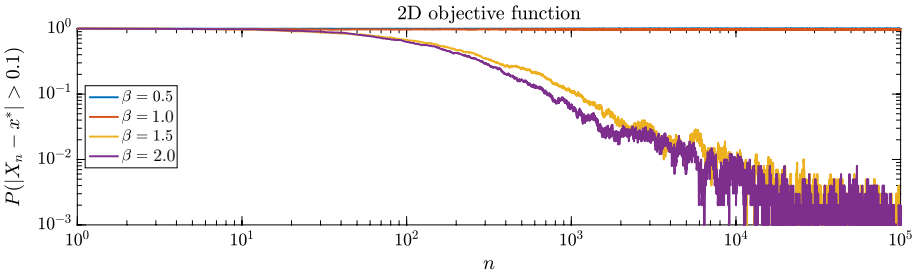
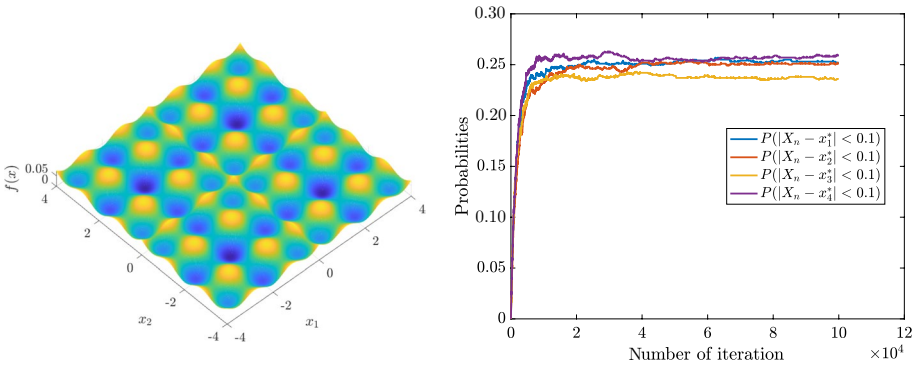


Fig. 8 Convergence history for minimizing (32) with $\varepsilon = 0$ and $d = 2$ after 10^5 number of iterations



(a) Optimization landscape with 4 global minimizers

(b) Convergence performance with $\varepsilon = 0$

Fig. 9 (**a**) An objective function landscape with four global minimizers labeled as $x_1^* = [2, 2]^\top$, $x_2^* = [-2, -2]^\top$, $x_3^* = [2, -2]^\top$, and $x_4^* = [-2, 2]^\top$; (**b**) convergence performance (from 10^3 i.i.d. runs) of the proposed algorithm with $\varepsilon = 0$, $\beta = 4$, and $d = 2$ after 10^5 number of iterations

sense since we have finite spatial resolution when computing the distributions. To be more precise, we set

$$\sigma_n(X_n) = \sqrt{2\Big[(f(X_n) - f_{\min}^*)^+\Big]^\beta},$$

which is (2). In Fig. 8, we plot the convergence histories for $d = 2$ and β ranging from 0.5 to 2, and we assume $f_{\min}^* = f_{\min} = 0$ is known. The probabilities in the y-axis are estimated using 10^3 i.i.d. runs while the x axis is the number of iterations. The initial guess is uniformly sampled from the domain Ω. It is worth noting that when $\beta < 1 = \frac{d}{2}$, there is no guarantee for convergence in probability since $\lim_{\varepsilon \to 0} Z_{\vec{u}} < \infty$ as defined in (16).

Next, we consider another optimization problem with four different global minima, whose optimization landscape is seen in Fig. 9a. We denote the global minimizers by $x_1^* = [2, 2]^\top$, $x_2^* = [-2, -2]^\top$, $x_3^* = [2, -2]^\top$, and $x_4^* = [-2, 2]^\top$. We implement the same algorithm with $\varepsilon = 0$ and $\beta = 4$, assuming again $f_{\min} = 0$ is known. The convergence behavior is shown in Fig. 9b. We can see that there are equal probabilities of roughly 25% for the iterate X_n to be in a close neighborhood of any of the four global minima for n large enough.

The numerical experiment in Fig. 3 shows that eliminating the regularization term $\varepsilon > 0$ in the algorithm (33) gives a faster convergence rate than that with the regularization term, at least for that particular objective function f. This requires that $f_{\min} = f(\mathbf{x}_*)$ is known. With an unknown optimal objective function value, there is a clear risk of having the algorithm trapped in local minima; see Fig. 5b. The regularization-free method, i.e., $\varepsilon = 0$, works very well when the optimization landscape is convex or when $f(\mathbf{x}_*)$ is known.

Lack of theoretical understanding for the case of $\varepsilon = 0$. We currently have a minimal theoretical understanding of the $\varepsilon = 0$ algorithm due to the strong degeneracy of the diffusion coefficient D in this case. The proof from Sect. 2 does not apply here because of the lack of appropriate Poincaré inequality in the strongly degenerate case, i.e., $\beta \geqslant d/2$. What we observe in the simulations might be an effect of discretization in the computational algorithm.

The degenerate elliptic operator in (6) is a challenge discussed extensively in the PDE and SDE literature; see, for example [4, 10, 13, 14, 19] and references therein. The classical way of handling degeneracy is to regularize the problem with a parameter ε and then take $\varepsilon \to 0$ [29, 34]. This allows one to establish the existence, and sometimes uniqueness, of the solution in a finite time interval $(0, T]$ but does not generalize to the limiting case of $T \to \infty$. There are recent results based on weighted estimates for the problem in the absence of the regularization parameter ε, mainly for the case of weak degeneracy, that is, when the exponent β is sufficiently small (see, for instance, reference [14] for a more precise definition of weak and strong degeneracy) [13, 14, 19]. In most cases, the existence of solutions to the Fokker-Planck equation (7) can only be established in the one-dimensional (1D) case (again in specific weighted function spaces) for a finite time interval $(0, T)$.

Existing results in simplified settings. There are indeed some precise characterizations of the singular behavior of such degenerate problems in simplified (yet still difficult) scenarios where the particular forms of diffusion coefficients (such as $D = x(1 - x)$ on $(0, 1)$) are assumed, for instance, in the case where the point of degeneracy (that is, the global minimizer in our case) is on the boundary of the domain and appropriate boundary conditions are prescribed at the point of degeneracy; see for instance [4, 10] for the detailed analysis of the Wright-Fisher equation. To demonstrate how the specific structure of the problem plays a role in the theory, let us consider the 1D case of $f(x) = x^2$ and $\beta = 1$. We further simplify the problem by taking $\Omega = \mathbb{R}$. With all these simplifications, we have that $D = x^2$, and the Fokker-Planck equation (7) simplifies to

$$u_t = (x^2 u)_{xx}, \quad -\infty < x < +\infty. \tag{38}$$

If we introduce the new variable $v = x^2 u$, we can check that v solves

$$v_t = x^2 v_{xx}, \quad -\infty < x < +\infty.$$

Due to the degeneracy at $x = 0$, we have that $v(0, t) = 0$. Therefore, we can focus only on the positive axis. The equation for v can be written as

$$v_t = x^2 v_{xx}, \quad 0 < x < +\infty, \quad v(0, t) = 0. \tag{39}$$

Let us perform the change of variable $x = \mathrm{e}^y$, that is, $y = \log x$. Then it is easy to check that the interval $(0, +\infty)$ is mapped to $(-\infty, +\infty)$. The Fokker-Planck equation is now mapped into the following constant-coefficient form:

$$\widetilde{v}_t = \widetilde{v}_{yy} - \widetilde{v}_y, \quad \widetilde{v}(-\infty, t) = 0. \tag{40}$$

With the boundary conditions, this system has a non-localized stationary distribution. This leads to the fact that $\int_{y_L}^{y_R} \widetilde{v} dy \nrightarrow 0$ as $t \to \infty$ for any finite interval (y_L, y_R). Using the fact that $v = x^2 u$, we conclude that $\int_{x_L}^{x_R} u dx \leqslant c\, x_R^{-2} \to 0$ as $x_R \to +\infty$ for some c, where $x_L = e^{y_L}$ and $x_R = e^{y_R}$. This simple argument shows that for any $x_L > 0$, $\int_{x_L}^{\infty} u dx = 0$. Therefore, the mass of u concentrates in the region $(0, x_L)$. This heuristic argument can be made more rigorous to show the concentration of the stationary distribution and can be generalized to the two-dimensional (2D) case with the radial function $f(x) = |x|^2$. Going beyond such specific forms seems extremely difficult.

3.3 Adding Gradient Information

One important goal of this paper is to prove that global convergence is possible with an algebraic rate without even approximating the gradient in the algorithm. Another goal is to develop an efficient derivative-free algorithm. Derivative-free methods typically compare different objective function values to find the direction for the next step or to accept a step or not. This is so for deterministic techniques, for example, the simplex method [24] and also for stochastic algorithms, for example, SA [22], and consensus-based optimization methods [2, 35].

Even if the gradient information is not necessary for convergence, adding such information from objective function values of several steps is also possible here. Without extra computational cost, the practical performance can be improved. We propose the following simple algorithm. First, we can accelerate the convergence with an approximated gradient based on the secant method as follows:

$$\overline{G}(X_n) = \sum_{i=1}^{I} w_i \frac{f(X_{n-i+1}) - f(X_{n-i})}{|X_{n-i+1} - X_{n-i}|^2} \left(X_{n-i+1} - X_{n-i}\right), \tag{41}$$

where $\sum_{i=1}^{I} w_i = 1$, and $w_i \geqslant 0$. For example, we can set the weight $w_i \sim \gamma^i$ for some $0 < \gamma < 1$. Using $\overline{G}(X_n)$ in place of the gradient term in a standard stochastic gradient descent scheme, we derive a modified algorithm compared to (1):

$$X_{n+1} = X_n - \eta_g \overline{G}(X_n) + \eta\, \sigma(f(X_n))\boldsymbol{\zeta}_n, \tag{42}$$

where η_g is the step size for the gradient term and other symbols follow earlier notations in (1).

We performed simulations using this algorithm with an estimated gradient. In Fig. 1, we present the result for the case when $f_{\min} = f(\mathbf{x}_*)$ is known a priori and $\sigma(X_n) = \sqrt{2(f(X_n) - f_{\min})^2}$ (that is, the case of $\beta = 2$ and $\varepsilon = 0$). We use the weights $w_i \sim \gamma^i$ where $\gamma = 0.5$ and various I values as used in (41). We compare the convergence performance of descent algorithms based on (1) and (42). The statistics are estimated from 10^3 i.i.d. runs. It is evident from the log-log plots in Fig. 1 that the approximated gradient information significantly accelerates the convergence of the stochastic descent algorithm when n is large. The semilog plots in Fig. 10 illustrate the *exponential* convergence when approximated gradients are used in the descent algorithm.

Next, we show an example of full-waveform inversion (FWI). FWI is a nonlinear inverse technique that utilizes the entire wavefield information to estimate the medium properties of the propagating domain. Without the loss of generality, the PDE constraint of FWI is the

following acoustic wave equation with zero initial condition and non-reflecting boundary conditions:

$$\begin{cases} m(\mathbf{x})\dfrac{\partial^2 u(\mathbf{x},t)}{\partial t^2} - \Delta u(\mathbf{x},t) = s(\mathbf{x},t), \\ u(\mathbf{x},0) = 0, \\ \dfrac{\partial u}{\partial t}(\mathbf{x},0) = 0. \end{cases} \tag{43}$$

We set the model parameter $m(\mathbf{x}) = 1/c(\mathbf{x})^2$, where $c(\mathbf{x})$ is the wave velocity, $u(\mathbf{x},t)$ is the forward wavefield, and $s(\mathbf{x},t)$ is the wave source. The velocity parameter m is often the target of reconstruction. Equation (43) is a linear PDE but defines a nonlinear operator $\mathcal{F}$ that maps $m(\mathbf{x})$ to $u(\mathbf{x},t)$. In FWI, we translate the inverse problem of finding the model parameter m based on the observable seismic data $\{g_i^{\text{obs}}\}$ to a constrained optimization problem:

$$m^* = \underset{m}{\operatorname{argmin}} f(m), \quad f(m) = \frac{1}{2}\sum_{i=1}^{n_s}\int_\Gamma\int_0^T \|g_i(x,t;m) - g_i^{\text{obs}}(x,t)\|^2 \mathrm{d}t\,\mathrm{d}x, \tag{44}$$

where n_s is the number of wave sources. For each given source $s_i(\mathbf{x},t)$ where $1 \leqslant i \leqslant n_s$, $g_i(x,t;m) = R\mathcal{F}(m)$ is the synthetic data with R being the linear projection operator that extracts the wavefield u_i at the measurement domain Γ.

We comment that (44) is a highly-nonconvex optimization problem. We will apply our AdaVar algorithm with an additional approximated gradient component (42) to find the global minimizer. First, we parameterize the velocity $c(\mathbf{x})$ to be piecewise-constant and we wish to invert ten unknowns $\{v_i\}_{i=1}^{10}$; see Fig. 11a for an illustration. That is, we search for $X = [v_1, \cdots, v_{10}] \in [1.5, 5.5]^{10} \subset \mathbb{R}^{10}$. Thus, the objective function can be denoted as $f(m) = f(v_1, \cdots, v_{10}) = f(X)$. In executing the algorithm (42), we set $\sigma(f(X_n)) = |f(X_n)|^3$, i.e., $\beta = 6 > d/2 = 5$, $\eta = 0.125$, $\eta_g = 0.05$, $\gamma = 0.5$, and $L = 2$, as the hyper-parameters. We consider $f_{\min} = 0$ since this is a data-fitting problem. Since the last layer right above the bottom boundary cannot be accurately recovered due to the non-reflective boundary condition, we assume its velocity is known to be 5 km/s. We place 8 sources and 60 receivers equally distributed on the top boundary. The source consists of two Ricker wavelets of disjoint supports at 15 Hz peak frequency. The ground truth is $X^* = [4.81, 4.77, 4.75, 4.83, 4.94, 5.35, 4.67, 4.83, 5.05, 5.18]$.

In Figs. 11b, c, we plot the convergence histories of the objective function values and the iterates. In the first 500 iterations, the update is dominated by noise as the objective function value, and the errors in the iterates fluctuate randomly. Later, the approximated

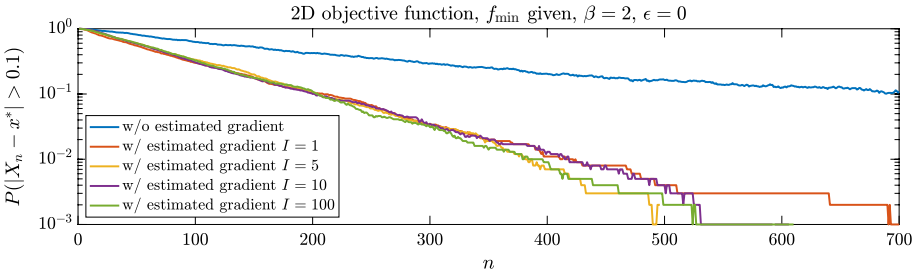
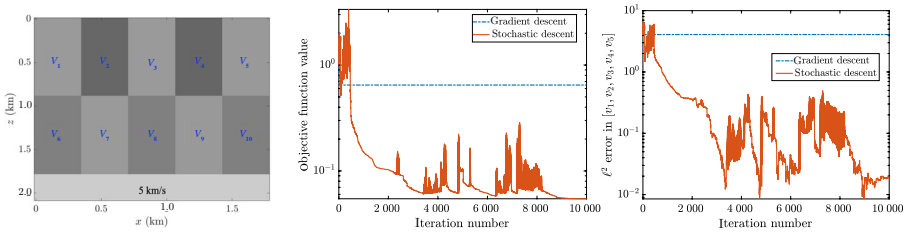


Fig. 10 Semilog plots of the same convergence statistics in Fig. 1 but for $n \leqslant 700$



(a) Velocity parameterization (b) Objective function value (c) ℓ^2 error in iterates

Fig. 11 Global optimization for FWI: the velocity parameterization with 10 unknowns (left), the objective function value decay (middle), and the convergence history of $[v_1, \cdots, v_5]$ (right) using the proposed stochastic algorithm and the standard gradient descent algorithm

gradient becomes the leading driving force since the objective function decays almost monotonically. The top-layer coefficients, $v_1, \cdots, v_5$, converge to the ground truth as measured in the ℓ^2 error; see Fig. 11c. The bottom-layer coefficients converge much slower as the objective function is not very sensitive to their changes, and the estimated gradient biases towards sensitive coefficients. We also plot the results using the gradient descent algorithm starting from a homogeneous velocity of 2 km/s. The iterates get stuck at a local minimum in fewer than 20 iterations, as we can see from Figs. 11b, c.

This method improves the convergence rate over (1) significantly, particularly in higher dimensions, as seen in the numerical experiments above. The original algorithm (1) does not suffer from the curse of dimensionality in the same way as in standard quadrature and PDE methods for which the discretization is done dimension by dimension. The algorithm here still shows severe degradation in modestly higher dimensions because β in (2) depends on d in determining the noise power $\sigma(f)$. The convergence of the classical gradient descent method is essentially independent of dimensional degradation. Thus, it is natural to add gradient information such as (42) to have a practical algorithm.

Remark 3 Here, we comment that the choice of β differs from our earlier discussions when the (approximated) gradient is present. The β values that give the best convergence for our derivative-free method are quite large; see Theorem 1. With explicitly adding the (approximated) gradient, the stochastic term with a large β is then too weak (given the fact that $f(x) - f^*_{\min} \in [0, 1]$ in our test cases) to escape a local minimum and overcome the adverse gradient in a reasonable time. This is particularly the case when the objective function value at the local minimum is close to the estimated optimum value $f^*_{\min}$. A smaller β naturally implies more noise based on the standard deviation $\sigma \sim |f(x) - f^*_{\min}|^{\beta/2}$ when $|f(x) - f^*_{\min}| \leqslant 1$, thereby increasing the probability of escape. Without the (approximated) gradient, the condition to ensure convergence in probability is that $\beta > d/2$; see Corollary 1. The same condition does not carry over to the case when the (approximated) gradient is present.

We can think of our estimated gradient $\overline{G}$ as a noisy version of the true gradient, i.e., $\overline{G} \approx \nabla f + \xi_k$. As a result, the iteration (42) with a constant σ converges, in probability, to the global minimizer of f under the right scaling (which essentially is that $\eta_g \sim 1/k$ and $\eta \sim 1/(k \log \log k)$. This can be shown with a slight modification of the techniques from [16], with minor additional assumptions on f.

4 Revisiting the AdaVar Stochastic Gradient Descent Algorithm in [9]

The basic concept of adding a stochastic term in the optimization algorithm with the variance of that term being state-dependent was already introduced in [9] for global optimization. The main algorithm proposed therein is discrete and of the form

$$X_{n+1} = X_n - \eta_n G(X_n) + \sigma_n\big(f(X_n)\big)\psi_n, \tag{45}$$

where $\{X_n\}$ are the iterates, $G(X_n)$ is the gradient of the objective function $f(x)$ evaluated at X_n, $\{\psi_n\}$ are i.i.d. samples from the standard d-dimensional Gaussian, and η_n is the step size. The noise power is controlled by the time- and state-dependent standard deviation σ_n defined as

$$\sigma_n(f(X_n)) = \begin{cases} \sigma_n^-, & f(X_n) \leqslant f_n, \\ \sigma_n^+, & f(X_n) > f_n, \end{cases} \tag{46}$$

where the sequence of scalar values $\{f_n\}$ is chosen by the user to implement the algorithm. Under proper control for the decay of f_n, and σ_n^-, together with further assumptions on all the other parameters, the results in [9] proved convergence of (45) to the global minimum of the objective function $f(x)$ at an algebraic rate in both probability and error in the state space. This was a significant improvement from the classic $\mathcal{O}(1/\sqrt{\log n})$ convergence using only time-dependent diffusion, such that given in [17].

The entire proof of the result in [9] is based on the discrete algorithm and its Markov property without using PDE or SDE. There are two main components to prove convergence, referred to as "Property One" and "Property Two" therein. The former is to show

$$\lim_{N\to\infty} \mathbb{P}\left(\bigcup_{n=0}^{N}\{X_n \in \Omega\}\right) = 1 \tag{47}$$

for any set Ω of the domain with a nonzero Lebesgue measure. This property ensures that the sequence of iterates generated by (45) will visit anywhere of the domain almost surely, thus including the global basin of attraction of the global minimizer. "Property Two", on the other hand, focuses on the local convergence. Once an iterate X_n lands in the global basin of attraction (i.e., a convex subset where the global minimizer belongs and the objective function is strictly convex), the probability that X_n stays in is much bigger than the probability of leaving the basin. As the sublevel set $\Omega_n = \{\mathbf{x}: f(\mathbf{x}) \leqslant f_n\}$ shrinks with f_n decreasing in (46), the "leaving" probability $\mathbb{P}(X_{n+1} \notin \Omega_n | X_n \in \Omega_n)$ might be large as the volume of Ω_n decreases while the "entering" probability $\mathbb{P}(X_{n+1} \in \Omega_n | X_n \notin \Omega_n)$ becomes smaller again due to the decreasing volume of Ω_n. The fact that the ratio between these two conditional probabilities goes to zero was key in the proof of [9] for "Property Two" and was driven by the gradient term. As one can see intuitively, the gradient information in (45) does not help with the iterates visiting everywhere of the domain, but it becomes extremely crucial to keep the iterate staying within the global basin. These two properties are the main building blocks, and we refer to [9, Sects. 3.1–3.2] for further details.

4.1 Differences

There are two main differences between our earlier paper [9] and this work. One is that the earlier algorithm explicitly included the gradient (see (45)), and the main proposal

in this paper is derivative-free (see (1)). In this paper, the step from X_n to X_{n+1} is taken at a uniformly random angle (due to the isotropic Gaussian noise in (1)). The optimization landscape must be explored in a sequence of many steps to find a descent direction, and our analysis in this work can, therefore, not be done in a Markovian way on the discrete level based on worst-case scenarios, which was done in [9]. We need the probability distribution of X_n here. Therefore, it is natural to study the continuum limit using the Fokker-Planck equation (7). This is common in convergence analysis; see, for example, [17].

We also comment that the convergence proof of [9] will not work without the gradient. The proof of "Property One", i.e., (47), will not be affected [9, Sect. 3.1]. This is because the arguments involved did not use the gradient explicitly, which, however, is essential in the proof of "Property Two" [9, Sect. 3.2], the local convergence. Without the gradient, the worst-case scenario in [9] will generate a very high probability for the iterate X_n to escape the set $\Omega_n = \{\mathbf{x} \in \Omega : f(\mathbf{x}) \leqslant f_n\}$, which is endowed with small noise variance; see (46). The discrete Markovian-style analysis will then not work. We have to use the history of X_n -values and the related probability density function to show that such worst-case scenarios have a small probability.

The other difference is the choice of the adaptive state-dependent noise term $\sigma(f)$. In [9], it was a piecewise-constant step function based on the value $f(\mathbf{x})$ as shown in (46), where f_n is a cut-off function decaying in n towards $f(\mathbf{x}_*)$. This was useful in its simplicity both for the analysis and in producing practical convergence. Without gradient information, noisy iterates driven by a constant variance will have a long hitting time to reach a close neighborhood of a global minimum. See the comments above and also in [9] for the importance of the gradient in this phase of the algorithm, i.e., "Property Two". Using a $\sigma(f)$, which is a strictly monotone function of the objective function value $f(\mathbf{x})$, will implicitly exploit gradient information over a sequence of steps throughout the full domain Ω. In this work, we indeed use such a regular monotone function of $|f(\mathbf{x}) - f^*_{\min}|$; see (2). The particular choice of a regular variance function in this paper fits nicely into the analysis of the Fokker-Planck equation.

From a practical point of view, the advantage of the currently proposed variance is the implicit encoding of the gradient, as remarked on in (9). The high-dimensional example in the earlier paper [9, Sect. 5.1.2] showed a much faster convergence into the basin of attraction of the global minimum than the result that a uniform sampling would have given. The gradient was a key in guiding the sequence of X_n-values closer to the optimum $\mathbf{x}_*$. If a step function-type variance is used in the derivative-free setup, we will only rely on uniform sampling to find the domain close to the optimum, resulting in very slow convergence.

4.2 Numerical Comparison

It is natural to ask how a monotone $\sigma(f)$ will do in the gradient descent algorithm of [9] and how the piecewise-constant variance (46) works without gradients in the framework of (1). The numerical examples below will shed light on these questions.

For numerical comparisons between the step function-based variance (46) proposed in [9] and the continuous variance (2) proposed in this work, we will first show when the gradient is present (that is, the approximated gradient in (41) is replaced by the real one), how the two strategies perform in global optimization. We use the 2D test case in [9, Eq.

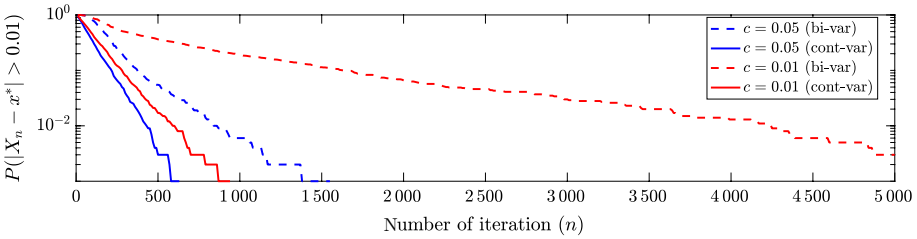


Fig. 12 Semilog-y plots of convergence performance for stochastic gradient descent with step function-based variance (46) and the continuous variance (2). The statistics are estimated by 10^3 i.i.d. runs

(5.1)]. Consider a variation of the so-called Rastrigin function [32] as our objective function $J_1(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \cdots, x_d)^\top \in \mathbb{R}^d$,

$$f_1(\mathbf{x}) = a\left(d - \sum_{i=1}^{d} \cos(bx_1)\right) + c\sum_{i=1}^{d} x_i^2. \tag{48}$$

When $c = 0$, all local minima of f_1 are global minima. When $c > 0$, f_1 has a unique global minimizer $x^* = (0, 0)$ and infinitely many local minima and saddle points in $\mathbb{R}^d$. We set the search domain Ω to be $[-20, 20]^d$, and consider $d = 2$, $a = b = 1$. In Fig. 12, we observe that for both $c = 0.01$ and $c = 0.05$, iterates driven by the continuous variance (labeled as "cont-var") converge much faster than the step function-based variance (labeled as "bi-var") in the plot. We applied the same strategy in choosing the hyper-parameters in the SGD algorithm with the step function-based variance (46) as in [9].

On the other hand, we can remove the gradient component in the SGD algorithm (45) from [9], leaving only the zero-mean noise with the two-stage variance (46) controlling the trajectory of the iterate X_n. As mentioned earlier, the same proof [9] will not go through. We can also observe the difficulty in convergence from the following numerical tests. Next, we consider the objective function (32) in two dimensions. The trajectory of one run using the continuous variance was shown earlier in Fig. 5a. Similarly, we plot the trajectory of

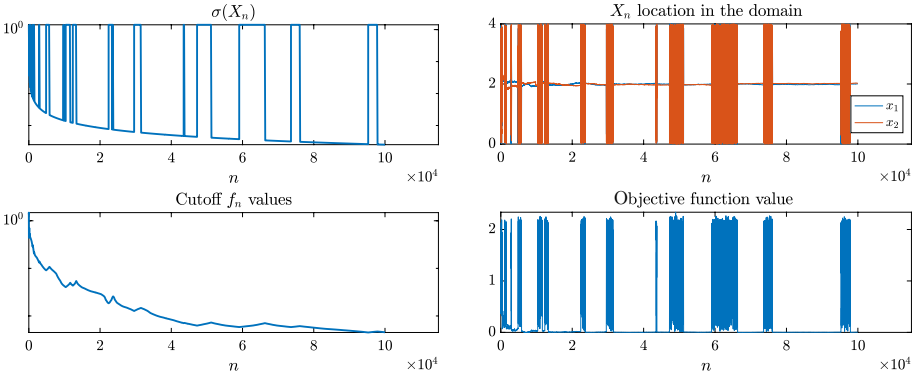


Fig. 13 Stochastic descent without gradient driven by the two-stage variance (46) for the 2D objective function (32). Top left: effective variance $\sigma(X_n)$ at the n-th iteration; bottom left: the cut-off value f_n in (46); top right: locations of X_n; bottom right: the objective function value $f(X_n)$

one run with the two-stage variance (46) in Fig. 13. The iterate X_n has been close to the global minimum $\mathbf{x}_* = [2, 2]^\top$ many times in the trajectory history but has also escaped shortly after.

5 Concluding Remarks

We have presented analysis and computational evidence to demonstrate the efficiency of an adaptive variance selection scheme for derivative-free optimization. While our theoretical justification is in the asymptotic regime, numerical simulations with the discrete algorithm show that the method works remarkably well in more challenging settings, for example, when the true value of the global minimum of the objective function is unknown.

The main difference between this contribution and other derivative-free methods is the rigorous analysis of global convergence with the algebraic rate, even in the case of no explicit gradient approximation. There are also several differences between the current work and our previous paper [9] in which a discrete version was studied. First, in paper [9], the proof of Property Two does not work without gradient information. Second, the probability distribution is needed in this work, and it depends on the objective function value of the iterate. The proof in [9] is instead based on the discrete algorithm. Several interesting theoretical issues remain to be addressed, including the convergence of the algorithm in the case of $\epsilon = 0$, having $\Omega = \mathbb{R}^d$, and including the estimation of $f_{\min}$ in the analysis. There are also more practical issues as, for example, the best choice of gradient approximation and involving parallel sequences of X_n in the optimization. We leave those to future works.

Appendix A: Weighted Poincaré Inequality

One key component in our analysis is the weighted Poincaré inequality with a given weight function $w(\mathbf{x})\colon \Omega \mapsto [0, \infty)$. To increase the readability of our proof, we recall here the inequality. The material here is standard and can be found in the references cited. For a general weight function $w(\mathbf{x})$, we first introduce the Muckenhoupt A_p weights.

Definition A1 (A_p *weights*) For a fixed $1 < p < \infty$, we say that a weight function $w\colon \mathbb{R}^d \mapsto [0, \infty)$ belongs to the class A_p if w is locally integrable, and for all cubes $Q \subset \mathbb{R}^d$, we have

$$[w]_p := \sup_{Q \subset \mathbb{R}^d} \left(\frac{1}{V_Q} \int_Q w(\mathbf{x})\, \mathrm{d}\mathbf{x} \right) \left(\frac{1}{V_Q} \int_Q w(\mathbf{x})^{-\frac{q}{p}}\, \mathrm{d}\mathbf{x} \right)^{\frac{p}{q}} < \infty, \tag{A1}$$

where q is a real number such that $\frac{1}{p} + \frac{1}{q} = 1$, and V_Q is the volume of the cube Q.

The following weighted Poincaré inequality for weights in the A_p class can be found in [30, Proposition 11.7].

Theorem A1 (Weighted Poincaré inequality [30]) *Let w be an A_p weight function and $f(\mathbf{x})$ a Lipschitz function. Then the following weighted Poincaré inequality holds for the hypercube $\Omega \subset \mathbb{R}^d$:*

$$\frac{1}{w(\Omega)} \int_\Omega |f - f_{\Omega,w}|^p w \, \mathrm{d}\mathbf{x} \leqslant \frac{2^p}{w(\Omega)} \int_\Omega |f - f_\Omega|^p w \, \mathrm{d}x \leqslant \frac{C_d^p \ell_\Omega^p [w]_p}{w(\Omega)} \int_\Omega |\nabla f|^p w \, \mathrm{d}\mathbf{x}, \tag{A2}$$

where $w(\Omega) = \int_\Omega w(\mathbf{x})\mathrm{d}\mathbf{x}$, $f_{\Omega,w} = \frac{1}{w(\Omega)} \int_\Omega f(\mathbf{x}) w(\mathbf{x})\mathrm{d}\mathbf{x}$, $f_\Omega = \frac{1}{V_\Omega} \int_\Omega f(\mathbf{x})\mathrm{d}\mathbf{x}$, ℓ_Ω *is the side length of the cube* Ω, *and* C_d *is a dimensional constant.*

There have been many results on the weighted Poincaré inequality [11, 19]. The paper by Pérez and Rela [30] improved some of the classical results and produced a quantitative control of the Poincaré constant (see (A1)) in the inequality, which is crucial for the analysis of our algorithm. We refer interested readers to [19, 30] for more general weighted Poincaré and Poincaré-Sobolev inequalities in various settings.

Remark A1 The definition of the A_p class allows one to consider degenerate and singular weights. For example, let $w(\mathbf{x}) = |\mathbf{x}|^\eta$, $\mathbf{x} \in \mathbb{R}^d$. Then $w \in A_p$ if and only if $-d < \eta < d(p-1)$.

Acknowledgements We are grateful for the valuable discussions with Professor Linan Chen (McGill University) and Professor Panagiotis E. Souganidis (University of Chicago) for constructive discussions. This work is partially supported by the National Science Foundation through grants DMS-2208504 (BE), DMS-1913309 (KR), DMS-1937254 (KR), and DMS-1913129 (YY). YY acknowledges support from Dr. Max Rössler, the Walter Haefner Foundation, and the ETH Zürich Foundation.

Data Availability The data that support the findings of this study are available upon reasonable request from the authors.

Compliance with Ethical Standards

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Alarie, S., Audet, C., Gheribi, A.E., Kokkolaras, M., Le Digabel, S.: Two decades of blackbox optimization applications. EURO J. Comput. Optim. **9**, 100011 (2021)
2. Carrillo, J.A., Jin, S., Li, L., Zhu, Y.: A consensus-based global optimization method for high dimensional machine learning problems. ESAIM Control Optim. Calc. Var. **27**, S5 (2021)
3. Cartis, C., Roberts, L.: Scalable subspace methods for derivative-free nonlinear least-squares optimization. Math. Program. **199**, 461–524 (2023)
4. Chen, L., Stroock, D.W.: The fundamental solution to the Wright-Fisher equation. SIAM J. Math. Anal. **42**, 539–567 (2010)
5. Chiang, T.-S., Hwang, C.-R., Sheu, S.J.: Diffusion for global optimization in $\mathbb{R}^n$. SIAM J. Control. Optim. **25**(3), 737–753 (1987)
6. Chow, S.-N., Yang, T.-S., Zhou, H.-M.: Global optimizations by intermittent diffusion. In: Chaos, CNN, Memristors and Beyond: a Festschrift for Leon Chua With DVD-ROM, composed by Eleonora Bilotta, pp. 466–479. World Scientific (2013)

7. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-free Optimization. SIAM, Philadelphia (2009)
8. Dekkers, A., Aarts, E.: Global optimization and simulated annealing. Math. Program. **50**(1), 367–393 (1991)
9. Engquist, B., Ren, K., Yang, Y.: An algebraically converging stochastic gradient descent algorithm for global optimization. arXiv:2204.05923 (2022)
10. Epstein, C.L., Mazzeo, R.: Wright-Fisher diffusion in one dimension. SIAM J. Math. Anal. **42**, 568–608 (2010)
11. Fabes, E.B., Kenig, C.E., Serapioni, R.P.: The local regularity of solutions of degenerate elliptic equations. Commun. Stat. Theory Methods **7**(1), 77–116 (1982)
12. Fornasier, M., Klock, T., Riedl, K:. Consensus-based optimization methods converge globally. arXiv:2103.15130v4 (2021)
13. Fragnelli, G., Mugnai, D.: Carleman estimates, observability inequalities and null controllability for interior degenerate non smooth parabolic equations. Memoirs of the American Mathematical Society, **242**(1146) (2016)
14. Fragnelli, G., Ruiz Goldstein, G., Goldstein, J.A., Romanelli, S.: Generators with interior degeneracy on spaces of l^2 type. Electron. J. Differ. Equ. **2012**, 1–30 (2012)
15. Frederick, C., Egerstedt, M., Zhou, H.: Collective motion planning for a group of robots using intermittent diffusion. J. Sci. Comput. **90**(1), 1–20 (2022)
16. Gelfand, S.B., Mitter, S.K.: Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. SIAM J. Control. Optim. **29**, 999–1018 (1991)
17. Geman, S., Hwang, C.-R.: Diffusions for global optimization. SIAM J. Control. Optim. **24**, 1031–1043 (1986)
18. Haupt, R.L., Haupt, S.E.: Practical Genetic Algorithms. Wiley, New York (2004)
19. Heinonen, J., Kipelainen, T., Martio, O.: Nonlinear Potential Theory of Degenerate Elliptic Equations. Courier Dover Publications, Mineola (2018)
20. Henderson, D., Jacobson, S.H., Johnson, A.W.: The theory and practice of simulated annealing. In: Handbook of Metaheuristics, pp. 287–319. Springer (2003)
21. Holland, J.H.: Holland. Genetic algorithms. Sci. Am. **267**, 66–73 (1992)
22. Kirkpatrick, S., Daniel Gelatt, C., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
23. Kolda, T.G., Lewis, R.M., Torczon, V.: Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev. **45**, 385–482 (2003)
24. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM J. Optim. **9**, 112–147 (1998)
25. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numer. **28**, 287–404 (2019)
26. McKinnon, K.I.M.: Convergence of the Nelder-Mead simplex method to a non-stationary point. SIAM J. Optim. **9**, 148–158 (1999)
27. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1998)
28. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**, 308–313 (1965)
29. Oleınik, O.A.: Linear equations of second order with nonnegative characteristic form. Mat. Sb. (NS) **69**, 111–140 (1966)
30. Pérez, C., Rela, E.: Degenerate Poincaré-Sobolev inequalities. Trans. Am. Math. Soc. **372**(9), 6087–6133 (2019)
31. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intell. **1**, 33–57 (2007)
32. Rastrigin, L.A.: Systems of Extremal Control. Nauka, Moscow (1974)
33. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: 1998 IEEE International Conference on Evolutionary Computation Proceedings, pp. 69–73 (1998)
34. Stroock, D.W., Srinivasa Varadhan, S.R.: Multidimensional Diffusion Processes. Springer Science & Business Media, Berlin (1997)
35. Totzeck, C.: Trends in consensus-based optimization. arXiv:2104.01383 (2021)