DISCRETE-TIME SIMULATED ANNEALING: A CONVERGENCE ANALYSIS VIA THE EYRING–KRAMERS LAW

WENPIN TANG, YUHANG WU, AND XUN YU ZHOU

Abstract. We study the convergence rate of the discrete-time simulated annealing process $(x_k;\, k = 0, 1, \ldots)$ for approximating the global optimum of a given function f. We prove that the tail probability $\mathbb{P}(f(x_k) > \min f + \delta))$ decays polynomial in cumulative step size, and provide an explicit rate through a non-asymptotic bound in terms of the model parameters. Our argument applies the recent development on functional inequalities for the Gibbs measure at low temperatures – the Eyring–Kramers law. The result leads to a condition on the step size to ensure the convergence. Finally, we perform numerical experiments to corroborate our theoretical result.

Key words: Simulated annealing, convergence rate, discrete time, Euler discretization, Eyring–Kramers law, functional inequalities, overdamped Langevin equation.

1. Introduction

Simulated annealing (SA) includes a set of stochastic optimization methods, whose goal is to find the global minimum of a function $f : \mathbb{R}^d \to \mathbb{R}$, in particular when f is nonconvex. These methods have many applications in physics, operations research and computer science; see e.g. van Laarhoven and Aarts (1987); Koulamas et al. (1994); Delahaye et al. (2019). The stochastic version of SA, independently proposed by Kirkpatrick et al. (1983) and Cerny (1985), considers a stochastic process related to f which is subject to thermal noise. When simulating this process, one decreases the temperature slowly over time. When this is done right, the stochastic process escapes from saddle points and local optima, and converges to the global minimum of f with high probability.

In this paper, we study the convergence rate of the *discrete-time SA process* $(x_k;\, k = 0, 1, \ldots)$ defined by

$$x_{k+1} = x_k - \nabla f(x_k)\eta_k + \sqrt{2\tau_{\Theta_k}\eta_k}Z_k, \quad x_0 \overset{d}{=} \mu_0(dx), \tag{1}$$

where η_k is the step size at iteration k, $\Theta_k := \sum_{j\le k}\eta_j$ is the cumulative step size up to iteration k, τ_{Θ_k} is the cooling schedule at iteration k, $(Z_k;\, k = 0, 1, \ldots)$ are independent and identically distributed standard normal vectors, and $\mu_0(dx)$ is some initial distribution. The algorithm (1) can be regarded as the Euler–Maruyama discretization of the *continuous-time SA process*, or the following *SA adapted overdamped Langevin equation* (Geman and Hwang, 1986):

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\tau_t}\, dB_t, \quad X_0 \overset{d}{=} \mu_0(dx), \tag{2}$$

where $(B_t;\, t \ge 0)$ is a standard Brownian motion in $\mathbb{R}^d$. For $\tau_t \equiv \tau$ constant in time, the scheme (1) is known as the *unadjusted Langevin algorithm* (ULA) which approximates

Date: February 4, 2024.

the Gibbs measure $\nu_\tau(dx) \propto \exp(-f(x)/\tau)dx$. The ULA was introduced by Parisi (1981); Grenander and Miller (1994), and further studied by Roberts and Tweedie (1996); Dalalyan (2017); Durmus and Moulines (2017).

The goal of this paper is to study the decay in time of the tail probability of (1), i.e. the deviation bound

$$\mathbb{P}(f(x_k) > \min f + \delta),$$

under suitable conditions on the function f, the discretization scheme η_k, Θ_k, and the cooling schedule τ_{Θ_k}. There are two motivations for studying this problem.

- First, there are growing interests in the interplay between sampling and optimization (Raginsky et al., 2017; Ma et al., 2019, 2021). The idea is to approximate the global optimum in nonconvex problems via Langevin dynamics-based stochastic gradient descent (Raginsky et al., 2017; Chen et al., 2020; Wang and Wu, 2020; Gao et al., 2022), along with its variants using non-reversibility (Hu et al., 2020) and replica exchange (Chen et al., 2019; Dong and Tong, 2021). Specifically, one aims to approximate $\min f$ by $\mathbb{E}f(x_k^\tau)$ where $(x_k^\tau;\, k = 0, 1, \ldots)$ is the ULA with a small, *fixed* temperature parameter τ. A drawback of this approach is that one needs to simulate *multiple* (*many*) sample paths to estimate $\mathbb{E}f(x_k^\tau)$. By contrast, the advantage of using SA processes is that for a suitable choice of *time-dependent* τ_{Θ_k}, the process x_k converges *almost surely* to $\min f$ as $k \to \infty$. Thus, one only needs to simulate *one* sample path to approximate $\min f$.
- Second, there have been recent research efforts in developing various noisy gradient-based algorithms (Ge et al., 2015; Jin et al., 2017; Chen et al., 2020; Guo et al., 2020) aiming at escaping saddle points and finding a local minimum of f as a surrogate. While finding a local surrogate has been proved to be sufficient in many machine learning problems, global optimization is important in its own right with applications ranging from finding Nash equilibria in various games (Myerson, 1991) to curriculum learning (Bengio et al., 2009). Compared with the gradient-based methods, SA sets finding global minima as the priority, if at the cost of longer exploration time.

The main tool in our analysis is the Eyring–Kramers law, which is a set of functional inequalities for the Gibbs measure at low temperatures (see Section 3.1 for details). To study the convergence rate of the discrete-time SA, it will be helpful to understand the long time behavior of its continuous analogue. It is well known (Geman and Hwang, 1986; Chiang et al., 1987) that the correct order of τ_t for the process (2) to converge to the global minimum of f is $(\ln t)^{-1}$, and there is a phase transition related to the *critical depth* E_* of the function f:

(a) If $\limsup_{t\to\infty} \tau_t \ln t \le E$ with $E < E_*$, then $\limsup_{t\to\infty} \mathbb{P}(f(X_t) \le \min f + \delta) < 1$.
(b) If $E \le \liminf_{t\to\infty} \tau_t \ln t \le \limsup_{t\to\infty} \tau_t \ln t < \infty$ with $E > E_*$, then

$$\lim_{t\to\infty} \mathbb{P}(f(X_t) \le \min f + \delta) = 1.$$

The formal definition of the critical depth E_* will be given in Assumption 2; see also Figure 1 below for an illustration when f is a double-well function. Roughly speaking, E_* is the largest hill one needs to climb starting from a local minimum to the global minimum. We refer to Tang and Zhou (2023) for background and further references.

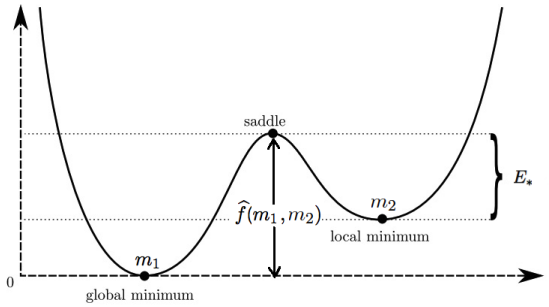


FIGURE 1. Illustration of the critical depth of a double-well function.

Building upon earlier works (Miclo, 1992; Menz and Schlichting, 2014; Menz et al., 2018), Tang and Zhou (2023) derive a non-asymptotic bound for the tail probability of the continuous-time SA (2) via a "four-step" analysis using the Eyring–Kramers law as a key technical tool. Their result is summarized as follows. To simplify the notation, we henceforth assume throughout this paper that

$$\min_{\mathbb{R}^d} f(x) = 0,$$

i.e. the global minimum of f is 0 by considering $f - \min f$.

Theorem A. *Assume that τ_t is decreasing in t, $\tau_t \sim \frac{E}{\ln t}$ with $E > E_*$, and $\frac{d}{dt}\left(\frac{1}{\tau_t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to \infty$. Then, under some assumptions on the function f, for any $\delta > 0$ there exists $C > 0$ independent of t such that*

$$\mathbb{P}(f(X_t) > \delta) \le C t^{-\min\left(\frac{\delta}{E}, \frac{1}{2}(1 - \frac{E_*}{E})\right)}.$$

Going back to the discrete-time SA process (1), a natural question is whether there is a similar convergence rate and under what additional conditions especially on the step size η_k. Our main result, which answers these questions, is outlined as follows. The precise statement of the result will be given in Section 2.

Theorem B. *Assume that τ_t is decreasing in t, $\tau_t \sim \frac{E}{\ln t}$ with $E > E_*$, and $\frac{d}{dt}\left(\frac{1}{\tau_t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to \infty$. Also assume that $\Theta_k \to \infty$ and $\limsup \eta_{k+1}\Theta_k < \infty$ as $k \to \infty$. Then, under some assumptions on the function f, for any $\delta > 0$ there exists $C > 0$ independent of t such that*

$$\mathbb{P}(f(x_k) > \delta) \le C \Theta_k^{-\min\left(\frac{\delta}{E}, \frac{1}{2}(1 - \frac{E_*}{E})\right)}.$$

This result of a non-asymptotic deviation bound for the discrete-time SA process is new to our best knowledge, and its proof is more involved and delicate than its continuous-time counterpart (Tang and Zhou, 2023) due to the discretization errors. It also gives a practical guidance on the choice of step size: the condition $\Theta_k \to \infty$ indicates that the step size cannot be chosen too small, while the condition $\limsup \eta_{k+1}\Theta_k < \infty$ suggests that the step size cannot be chosen too large. For instance, $\eta_k = k^{-\theta}$ with $\theta \in [\frac{1}{2}, 1]$ satisfies the conditions in the theorem to ensure the convergence. Also note that the rate $\min\left(\frac{\delta}{E}, \frac{1}{2}(1 - \frac{E_*}{E})\right)$ is smaller than $\frac{1}{2}$. Empirical results in Section 5 suggest that this rate is optimal, but it remains open to prove a matching lower bound. We leave the problem for future work.

The dependence of the constant C on the dimension d is another interesting problem. It is also a subtle problem, since most analysis including the Eyring–Kramers law uses Laplace's method, while the latter may fail if both the dimension d and the inverse temperature $1/\tau$ tend to infinity (Shun and McCullagh, 1995). As explained in (Tang and Zhou, 2023, Remark 1), an upper bound for C is exponential in d. This suggests the convergence rate is exponentially slow as the dimension increases, which aligns with the fact that finding the global minimum of a general nonconvex function is NP-hard.

Note that the discrete-time SA (1) belongs to the general class of stochastic gradient descent algorithms of the form:

$$x_{k+1} = x_k - a_k \nabla f(x_k) + b_k Z_k, \quad x_0 \overset{d}{=} \mu_0(dx),$$

where $(a_k;\, k = 0, 1, \ldots)$ and $(b_k;\, k = 0, 1, \ldots)$ are two positive deterministic sequences. Most of the existing literature (e.g. Raginsky et al., 2017; Chen et al., 2020) deals with ULA or its variants with a fixed (though small) temperature parameter τ, corresponding to the *strongly perturbed condition* where a_k/b_k^2 is assumed to be of a constant order. It is clear that the discrete SA process (1) does not satisfy the strongly perturbed condition under which the process x_k would converge in distribution to a diffuse measure instead of a Dirac mass. On the other hand, Gelfand and Mitter (1991) show that for $a_k = \frac{1}{k}$ and $b_k = \frac{b}{\sqrt{k \log\log k}}$ with some large $b > 0$, the process $f(x_k)$ converges in probability to $\min f$; however, they do not give any convergence rate. Pelletier (1998) proves, under the *annealing condition* that $a_k/(b_k^2 \ln k)$ is of a constant order, that $\frac{4a_k}{b_k^2}(f(x_k) - \min f)$ converges in distribution to a Gamma random variable. This corresponds to the central limit theorem or small deviation regime ($\delta = \delta_k \downarrow 0$), while in this paper we are concerned with large deviation regime (δ is fixed) which is more practically meaningful. Indeed, the special time-annealing nature of the perturbation term in the discrete SA process makes the problem more challenging, which is the reason why the Eyring–Kramers asymptotics in low temperatures is needed. For instance, the condition $\limsup \eta_{k+1}\Theta_k < \infty$ stems from the one-iteration estimate via the Eyring–Kramers formula.

The remainder of the paper is organized as follows. Section 2 presents the assumptions and our main result. Section 3 provides background on functional inequalities, and sketches the main idea in proving Theorem A for the convergence rate of the continuous-time SA process. The latter is useful for the reader to understand the main difficulty in extending the idea to the discrete-time case. The main result (Theorem 1) is proved in Section 4. Results of numerical experiments on global optimization are reported in Section 5. We conclude in Section 6.

2. Main result

In this section, we make precise the informal statement in the introduction by presenting the main result of the paper. We first collect the notations that will be used throughout this paper.

- $|\cdot|$ is the Euclidean norm of a vector, and $a \cdot b$ is the scalar product of vectors a and b.
- For a function $f : \mathbb{R}^d \to \mathbb{R}$, ∇f, $\nabla^2 f$ and Δf are its gradient, Hessian and Laplacian respectively.

– $a \sim b$ means that $a/b \to 1$ as some problem parameter tends to 0 or ∞. Similarly, $a = \mathcal{O}(b)$ means that a/b is bounded as some problem parameter tends to 0 or ∞.

We use C for a generic constant which depends on problem parameters $(\delta, f, E \ldots)$, whose values may change from line to line.

Next, we present a few assumptions on the function f. These assumptions are standard in the study of metastability; see Menz and Schlichting (2014); Menz et al. (2018).

Assumption 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be smooth, bounded from below, and satisfy the conditions:*

(i) f is non-degenerate on the set of critical points. That is, for some $C > 0$,

$$\frac{|\xi|}{C} \leq |\nabla^2 f(x)\xi| \leq C|\xi| \quad \text{for each } x \in \{z : \nabla f(z) = 0\} \text{ and } \xi \in \mathbb{R}^d.$$

(ii) There exists $C, C' > 0$ such that

$$\liminf_{|x|\to\infty} \frac{|\nabla f(x)|^2 - \Delta f(x)}{|x|^2} \geq C, \quad \inf_x \nabla^2 f(x) \geq -C'.$$

Let us make a few comments on Assumption 1. The condition (ii) is a version of the *dissipative condition*, and it implies that f has at least quadratic growth at infinity. This is a necessary and sufficient condition to obtain the log-Sobolev inequality (see Royer, 2007, Theorem 3.1.21) which is key to convergence analysis. The conditions (i) and (ii) imply that the set of critical points is discrete and finite (Menz and Schlichting, 2014, Remark 1.6). In particular, it follows that the set of local minimum points $\{m_1, \ldots, m_N\}$ is also finite, with N the number of local minimum points of f.

Define the saddle height $\widehat{f}(m_i, m_j)$ between two local minimum points m_i, m_j by

$$\widehat{f}(m_i, m_j) := \inf \left\{ \max_{s\in[0,1]} f(\gamma(s)) : \gamma \in \mathcal{C}[0,1],\, \gamma(0) = m_i,\, \gamma(1) = m_j \right\}. \tag{3}$$

See Figure 1 for an illustration of the saddle height $\widehat{f}(m_0, m_1)$ when f is a double-well function with m_0 the global minimum and m_1 the local minimum.

Assumption 2. *Let $m_1, \cdots, m_N$ be the positions of the local minima of f.*

(i) m_1 is the unique global minimum point of f, and $m_1, \ldots, m_N$ are ordered in the sense that there exists $\delta > 0$ such that

$$f(m_N) \geq f(m_{N-1}) \geq \cdots \geq f(m_2) \geq \delta \quad \text{and} \quad f(m_1) = 0.$$

(ii) For each $i, j \in \{1, \ldots, N\}$, the saddle height between m_i, m_j is attained at a unique critical point s_{ij} of index one. That is, $f(s_{ij}) = \widehat{f}(m_i, m_j)$, and if $\{\lambda_1, \ldots, \lambda_n\}$ are the eigenvalues of $\nabla^2 f(s_{ij})$, then $\lambda_1 < 0$ and $\lambda_i > 0$ for $i \in \{2, \ldots, n\}$. The point s_{ij} is called the communicating saddle point between the minima m_i and m_j.

(iii) There exists $p \in [N]$ such that the energy barrier $f(s_{p1}) - f(m_p)$ dominates all the others. That is, there exists $\delta > 0$ such that for all $i \in [N] \setminus \{p\}$,

$$E_* := f(s_{p1}) - f(m_p) \geq f(s_{i1}) - f(m_i) + \delta.$$

The dominating energy barrier E_* is called the critical depth.

The above two assumptions are also imposed in the continuous-time counterpart of Tang and Zhou (2023). To get the convergence result for the discrete-time simulated annealing, we need an additional condition on the function f.

Assumption 3. *The gradient ∇f is globally Lipschitz, i.e. $|\nabla f(x) - \nabla f(y)| \le L|x-y|$ for some $L > 0$.*

The convergence result for the discrete-time SA process (1) is stated as follows, whose proof is deferred to Section 4.

Theorem 1. *Let f satisfy Assumptions 1, 2 & 3, and let μ_0 satisfy the moment condition: for each $p \ge 1$, there exists $C_p > 0$ such that*

$$\int_{\mathbb{R}^d} f(x)^p \mu_0(dx) \le C_p. \tag{4}$$

Assume that τ_t is decreasing in t, $\tau_t \sim \frac{E}{\ln t}$ with $E > E_*$, and $\frac{d}{dt}\left(\frac{1}{\tau_t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to \infty$. Moreover, assume that $\Theta_k \to \infty$ and

$$\limsup \eta_{k+1}\Theta_k < \infty, \tag{5}$$

as $k \to \infty$. Then for each $\delta, \varepsilon > 0$, there exists $C > 0$ independent of t such that

$$\mathbb{P}(f(x_k) > \delta) \le C\Theta_k^{-\min\left(\frac{\delta}{E}, \frac{1}{2}(1-\frac{E_*}{E})\right)+\varepsilon}. \tag{6}$$

3. Preliminaries

In this section, we recall some basic results of functional inequalities and explain how these results are applied in the setting of SA. We also highlight the "four step" analysis in proving the continuous-time counterpart of Theorem 1, which sheds light on how we prove in the discrete setting and under the difference.

3.1. Functional inequalities and the Eyring-Kramers law.

Let ν_τ be the Gibbs measure with landscape $f(\cdot)$ and temperature τ defined by

$$\nu_\tau(dx) = \frac{1}{Z_\tau} \exp\left(-\frac{f(x)}{\tau}\right) dx, \tag{7}$$

where $Z_\tau := \int_{\mathbb{R}^d} \exp(-f(x)/\tau)dx$ is the normalizing constant. It is well known that under suitable conditions on f, $\nu_\tau(dx)$ is the stationary distribution of the overdamped Langevin equation

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\tau}\, dB_t, \quad X_0 \overset{d}{=} \mu_0(dx). \tag{8}$$

The difference between the overdamped Langevin process (8) and the continuous-time SA (2) is that the temperature τ_t of the latter is decreasing in time. Due to the time dependence, the limiting distribution of the solution to (2) is unknown. As we will see in Section 3.2, the idea is to approximate (2) by a process of Gibbs measures with temperature τ_t. Since τ_t decreases to 0 in the limit, the problem boils down to studying Gibbs measures at low temperatures.

Now we present functional inequalities of Gibbs measures at low temperatures ($\tau \to 0$). Let μ and ν be two probability measures on $\mathbb{R}^d$ such that μ is absolutely continuous relative

to ν, with $d\mu/d\nu$ being the Radon-Nikodym derivative. Define the relative entropy or KL-divergence $H(\mu|\nu)$ of μ with respect to ν by

$$H(\mu|\nu) := \int \log\left(\frac{d\mu}{d\nu}\right) d\mu = \int \frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) d\nu, \tag{9}$$

and the Fisher information $I(\mu|\nu)$ of μ with respect to ν by

$$I(\mu|\nu) := \frac{1}{2} \int \left|\nabla\left(\frac{d\mu}{d\nu}\right)\right|^2 \left(\frac{d\mu}{d\nu}\right)^{-1} d\nu. \tag{10}$$

We say that a probability measure ν satisfies the log-Sobolev inequality (LSI) with constant $\alpha > 0$, if for all probability measures μ with $I(\mu|\nu) < \infty$,

$$H(\mu|\nu) \le \frac{1}{\alpha} I(\mu|\nu). \tag{11}$$

The constant α is called the LSI constant for the probability measure ν. For instance, the LSI constant $\alpha = 1$ for ν the multivariate Gaussian with mean 0 and covariance matrix I_d.

Assume that the Gibbs measure ν_τ defined by (7) satisfies the LSI with constant $\alpha_\tau > 0$. The subscript 'τ' in α_τ suggests the dependence of the LSI constant on the temperature τ, and we are interested in the asymptotics of α_τ at low temperatures as $\tau \to 0$. This problem was considered by (Menz and Schlichting, 2014, Corollary 3.18), where a sharp lower bound for α_τ as $\tau \to 0$ was derived.

Lemma 1. *Let f satisfy Assumptions 1 & 2. Then the Gibbs measure ν_τ defined by (7) satisfies the LSI with constant $\alpha_\tau > 0$ such that*

$$\alpha_\tau \sim C \exp\left(-\frac{E_*}{\tau}\right) \quad \text{as } \tau \to 0, \tag{12}$$

where $C > 0$ depends on f, d.

The Eyring–Kramers law provides an estimate on the spectral gap of the overdamped Langevin equation (8). Lemma 1 is the LSI version of the Eyring–Kramers law, which is stronger than the spectral gap estimate implied by the Poincaré inequality (Bovier et al., 2004, 2005).

Define the Wasserstein distance $W_2(\mu, \nu)$ between μ and ν by

$$W_2(\mu, \nu) := \inf_{\Pi} \sqrt{\int |x - y|^2 \Pi(dx, dy)}, \tag{13}$$

where the infimum is over all joint distributions Π coupling μ and ν. We say that a probability measure ν satisfies Talagrand's inequality with constant $\gamma > 0$, if for all probability measure μ with $H(\mu|\nu) < \infty$,

$$W_2(\mu, \nu) \le \sqrt{\frac{2}{\gamma} H(\mu|\nu)}. \tag{14}$$

It follows from (Otto and Villani, 2000, Theorem 1) that the LSI implies Talagrand's inequality with the same constant, namely, if ν satisfies the LSI with constant $\alpha > 0$, then ν also satisfies Talagrand's inequality with constant $\gamma = \alpha$. Combining with Lemma 1, we get a lower bound estimate of Talagrand's inequality constant for the Gibbs measure ν_τ.

Lemma 2. *Let f satisfy Assumptions 1 & 2. Then the Gibbs measure ν_τ defined by (7) satisfies Talagrand's inequality with constant $\gamma_\tau > 0$ such that*

$$\gamma_\tau \sim C \exp\left(-\frac{E_*}{\tau}\right) \quad \text{as } \tau \to 0 \tag{15}$$

where $C > 0$ depends on f, d.

3.2. **Proof sketch of Theorem A.** Here we sketch the proof of Theorem A provided in Tang and Zhou (2023), which will help understand the proof techniques in Section 4.

Let μ_t be the probability measure of X_t defined by (2). The key idea is to compare μ_t with the time-dependent Gibbs measure ν_{τ_t} given by

$$\nu_{\tau_t}(dx) = \frac{1}{Z_{\tau_t}} \exp\left(-\frac{f(x)}{\tau_t}\right) dx,$$

where $Z_{\tau_t} := \int_{\mathbb{R}^d} \exp(-f(x)/\tau_t)$ is the normalizing constant. Note that ν_{τ_t} will concentrate on the minimum point of f as $t \to \infty$ since $\tau_t \to 0$ as $t \to \infty$. We will see that ν_{τ_t} is close to μ_t in some sense as $t \to \infty$. The proof of Theorem A is broken into four steps.

Step 1: Reduce μ_t to ν_{τ_t}. Let $(\widetilde{X}_t;\, t \geq 0)$ be a process whose distribution is ν_{τ_t} at time t. By a simple coupling argument and Pinsker's inequality, we have

$$\mathbb{P}(f(X_t) > \delta) \leq \mathbb{P}(f(\widetilde{X}_t) > \delta) + \sqrt{2H(\mu_t|\nu_{\tau_t})}. \tag{16}$$

So the problem boils down to estimating $\mathbb{P}(f(\widetilde{X}_t) > \delta)$ and $H(\mu_t|\nu_{\tau_t})$.

Step 2: Long-time behavior of $f(\widetilde{X}_t)$. Apply Laplace's method to show that for each $\varepsilon \in (0, \delta)$, there exist $C > 0$ independent of t such that

$$\mathbb{P}(f(\widetilde{X}_t) > \delta) \leq C t^{-\frac{\delta - \varepsilon}{E}}. \tag{17}$$

Step 3: Differential inequality for $H(\mu_t|\nu_{\tau_t})$. Apply the Fokker–Planck equation of the over-damped Langevin equation and integration by parts to show

$$\frac{d}{dt} H(\mu_t|\nu_{\tau_t}) \leq -2\tau_t I\left(\mu_t|\nu_{\tau_t}\right) + \frac{d}{dt}\left(\frac{1}{\tau_t}\right) \mathbb{E} f(X_t) \tag{18}$$

for any τ_t decreasing in t.

Step 4: Estimating $H(\mu_t|\nu_{\tau_t})$ via the Eyring–Kramers law. There are two terms on the right hand side of (18). It is easy to show that

$$\mathbb{E} f(X_t) \leq C(1+t)^\varepsilon. \tag{19}$$

Hence, by Lemma 1 and the inequalities (18), (19), we have

$$\begin{aligned} \frac{d}{dt} H(\mu_t|\nu_{\tau_t}) &\leq -2\tau_t \alpha_t H(\mu_t|\nu_{\tau_t}) + \frac{C}{t}\, \mathbb{E} f(X_t) \\ &\leq -C t^{-\left(\frac{E_*}{E} - \varepsilon\right)} H(\mu_t|\nu_{\tau_t}) + C t^{-1+\varepsilon}, \end{aligned}$$

where α_t is the LSI constant for the Gibbs measure ν_{τ_t}. By Grönwall's inequality, we get

$$H(\mu_t|\nu_{\tau_t}) \leq C t^{-1+\frac{E_*}{E}+2\varepsilon}. \tag{20}$$

Combining (16), (17) and (20) proves Theorem A.

4. Proof of Theorem 1

This section is devoted to the proof of Theorem 1. While the essential idea is built upon that employed for the continuous-time SA process sketched in Section 3.2, subtle additional analysis is called for due to discretization.

Recall that η_k is the step size at iteration k, and $\Theta_k = \sum_{j\le k}\eta_j$ is the cumulative step size up to iteration k. Let μ_k be the probability density of x_k defined by (1), and

$$\nu_{\tau_{\Theta_k}}(dx) = \frac{1}{Z_{\tau_{\Theta_k}}}\exp\left(-\frac{f(x)}{\tau_{\Theta_k}}\right)dx, \tag{21}$$

where $Z_{\tau_{\Theta_k}} := \int_{\mathbb{R}^d}\exp(-f(x)/\tau_{\Theta_k})dx$ is the normalizing constant. We divide the proof into four steps.

Step 1: Reduce μ_k to $\nu_{\tau_{\Theta_k}}$. This step is similar to Steps 1 & 2 for the continuous-time case described in Section 3.2. Let $(\widetilde{x}_k;\, k\ge 0)$ be a sequence whose distribution is $\nu_{\tau_{\Theta_k}}$ at epoch k, living on the same probability space as $(x_k;\, k\ge 0)$. Fix $\delta > 0$. The same argument as in (16) shows that

$$\mathbb{P}(f(x_k) > \delta) \le \mathbb{P}(f(\widetilde{x}_k) > \delta) + \sqrt{2H(\mu_k|\nu_{\tau_{\Theta_k}})}. \tag{22}$$

Assume that $\Theta_k \to \infty$ and $\tau_{\Theta_k} \sim \frac{E}{\ln\Theta_k}$ as $k\to\infty$ with $E > E_*$. Similar to (17), we get a bound for the first term on the right hand side of (22). That is, for each $\varepsilon \in (0,\delta)$, there exists $C > 0$ independent of t such that

$$\mathbb{P}(f(\widetilde{x}_k) > \delta) \le C\Theta_k^{-\frac{\delta-\varepsilon}{E}}. \tag{23}$$

So it remains to estimate $H(\mu_k|\nu_{\tau_{\Theta_k}})$, which is the task of the next three steps.

Step 2: Continuous-time coupling. To make use of continuous-time tools, we couple the sequence $(x_k;\, k\ge 0)$ by a continuous-time process $(X_t;\, t\ge 0)$ such that $(X_{\Theta_k};\, k\ge 0)$ has the same distribution as $(x_k;\, k\ge 0)$. To do this, define the process X by

$$dX_t = -\nabla f(x_k)dt + \sqrt{2\tau_{\Theta_k}}dB_t, \quad t\in[\Theta_k,\Theta_{k+1}), \tag{24}$$

where we identify X_{Θ_k} with x_k. So X on $[\Theta_k,\Theta_{k+1})$ is Brownian motion with drift $-\nabla f(x_k)$ and covariance $\sqrt{2\tau_{\Theta_k}}I_d$. As mentioned in Step 3, Section 3.2, the Fokker–Planck equation plays an important role in the analysis of the continuous-time SA process. It is desirable to get a version of the Fokker–Planck equation for the coupled process (24). The result is stated as follows.

Lemma 3. *For $t\in[\Theta_k,\Theta_{k+1})$, the probability density μ_t of X_t defined by* (24) *satisfies the following equation:*

$$\frac{\partial\mu_t}{\partial t} = \nabla\cdot\left(\tau_{\Theta_k}\nu_{\tau_{\Theta_k}}\nabla\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right)\right) + \nabla\cdot\left(\mu_t\,\mathbb{E}[\nabla f(x_k) - \nabla f(X_t)|X_t = x]\right). \tag{25}$$

Proof. Let $\mu_{t|s}(x|y)$ be the conditional probability $\mathbb{P}(X_t = x|X_s = y)$. By conditioning on $X_{\Theta_k} = x_k$, we have

$$\frac{\partial\mu_{t|\Theta_k}(x|x_k)}{\partial t} = \nabla\cdot(\mu_{t|\Theta_k}(x|x_k)\nabla f(x_k)) + \tau_{\Theta_k}\Delta\mu_{t|\Theta_k}(x|x_k). \tag{26}$$

Integrating (26) against μ_{Θ_k} and using the fact that $\mu_{t|\Theta_k}(x|x_k)\mu_{\Theta_k}(x_k) = \mu_t(x)\mu_{\Theta_k|t}(x_k|x)$, we get

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t(x)\, \mathbb{E}[\nabla f(x_k)|X_t = x]) + \tau_{\Theta_k} \Delta \mu_t. \tag{27}$$

Further by the Fokker–Planck equation of the overdamped Langevin equation, we have

$$\nabla \cdot \left(\tau_{\Theta_k} \nu_{\tau_{\Theta_k}} \nabla \left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}} \right) \right) = \nabla \cdot (\mu_t \nabla f(x)) + \tau_{\Theta_k} \Delta \mu_t. \tag{28}$$

Combining (27) and (28) yields (25). □

There are two terms on the right hand side of (25). The first term is the usual Fokker–Planck term, while the second term corresponds to the discretization error.

Step 3: One-step analysis of $H(\mu_k|\nu_{\tau_{\Theta_k}})$. Here we use the coupled process (24) to study the one-step decay of $H(\mu_k|\nu_{\tau_{\Theta_k}})$.

Lemma 4. *Let f satisfy Assumptions 1, 2 & 3, and assume that the condition (4) for μ_0 holds. Assume that τ_t is decreasing in t, $\tau_t \sim \frac{E}{\ln t}$ with $E > E_*$, and $\frac{d}{dt}\left(\frac{1}{\tau_t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to \infty$. Also assume that $\Theta_k \to \infty$ and $\eta_{k+1}\Theta_k \to 0$ as $k \to \infty$. Then, for each $\varepsilon > 0$, there exist $C, C' > 0$ independent of t such that*

$$\begin{aligned} H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) \le \left(1 - C\eta_{k+1}\Theta_k^{-(\frac{E_*}{E}+\varepsilon)}\right) H(\mu_k|\nu_{\tau_{\Theta_k}}) \\ + C'(\eta_{k+1}^2 + \eta_{k+1}^3 \ln \Theta_k + \eta_{k+1}\Theta_k^{-1+\varepsilon}). \end{aligned} \tag{29}$$

Proof. Write

$$H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) = \underbrace{H(\mu_{k+1}|\nu_{\tau_{\Theta_k}})}_{(a)} + \underbrace{\left(H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) - H(\mu_{k+1}|\nu_{\tau_{\Theta_k}})\right)}_{(b)}. \tag{30}$$

We first use the coupled process (24) to study the term (a). Note that

$$\begin{aligned} \frac{d}{dt} H(\mu_t|\nu_{\tau_{\Theta_k}}) &= \int \frac{\partial \mu_t}{\partial t} \ln\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right) dx + \int \mu_t \frac{d}{dt} \ln\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right) dx \\ &= \int \nabla \cdot \left(\tau_{\Theta_k} \nu_{\tau_{\Theta_k}} \nabla \left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}} \right) \right) \ln\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right) dx \\ &\quad + \underbrace{\int \nabla \cdot (\mu_t\, \mathbb{E}[\nabla f(x_k) - \nabla f(X_t)|X_t = x]) \ln\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right) dx}_{(c)} + \frac{d}{dt} \int \mu_t(dx) \\ &= -2\tau_{\Theta_k} I(\mu_t|\nu_{\tau_{\Theta_k}}) + (c), \end{aligned} \tag{31}$$

where we use (25) in the second equation, and the fact that $\int \nabla \cdot \left(\tau_t \nu_{\tau_t} \nabla \left(\frac{\mu_t}{\nu_{\tau_t}} \right) \right) \ln\left(\frac{\mu_t}{\nu_{\tau_t}}\right) dx = -2\tau_t I(\mu_t|\nu_{\tau_t})$ in the third equation. Now we need to estimate the term (c) in (31). By

integration by parts and the fact that $a \cdot b \leq \frac{1}{\tau_{\Theta_k}}|a|^2 + \frac{\tau_{\Theta_k}}{4}|b|^2$, we get

$$\begin{aligned}(c) &= \mathbb{E}\left((\nabla f(X_t) - \nabla f(x_k)) \cdot \nabla \ln\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right)\right) \\ &\leq \frac{1}{\tau_{\Theta_k}}\mathbb{E}|\nabla f(X_t) - \nabla f(x_k))|^2 + \frac{\tau_{\Theta_k}}{4}\mathbb{E}\left|\nabla \ln\left(\frac{\mu_t}{\nu_{\tau_{\Theta_k}}}\right)\right|^2 \\ &\leq \frac{L^2}{\tau_{\Theta_k}}\mathbb{E}|X_t - x_k|^2 + \frac{\tau_{\Theta_k}}{2} I(\mu_t|\nu_{\tau_{\Theta_k}}), \end{aligned} \tag{32}$$

where the expectation $\mathbb{E}$ is with respect to $\mu_t(dx)$, and L is the Lipschitz constant of ∇f in Assumption 3. Recall from (24) that $X_t - x_k = -\nabla f(x_k)(t - \Theta_k) + \sqrt{2\tau_{\Theta_k}(t-\Theta_k)}Z$, where Z is standard normal. Hence

$$\begin{aligned}\mathbb{E}|X_t - x_k|^2 &= (t-\Theta_k)^2\mathbb{E}|\nabla f(x_k)|^2 + 2\tau_{\Theta_k}(t-\Theta_k)d \\ &\leq \eta_{k+1}^2\mathbb{E}|\nabla f(x_k)|^2 + C\tau_{\Theta_k}\eta_{k+1}. \end{aligned} \tag{33}$$

According to Lemma 2, $\nu_{\tau_{\Theta_k}}$ satisfies Talagrand's inequality with constant $\gamma_{\tau_{\Theta_k}} \sim \kappa \exp(-E_*/\tau_{\Theta_k})$. Moreover, by (Vempala and Wibisono, 2019, Lemma 10),

$$\mathbb{E}|\nabla f(x_k)|^2 \leq \frac{C}{\gamma_{\tau_{\Theta_k}}} H(\mu_k|\nu_{\tau_{\Theta_k}}) + C. \tag{34}$$

Combining (32) with (33), (34) and the fact that $\tau_{\Theta_k} \sim \frac{E}{\ln \Theta_k}$ as $k \to \infty$, we have

$$(c) \leq C\left(\eta_{k+1}^2 \Theta_k^{\frac{E_*}{E}} \ln \Theta_k\right) H(\mu_k|\nu_{\tau_{\Theta_k}}) + C(\eta_{k+1} + \eta_{k+1}^2 \ln \Theta_k) + \frac{\tau_{\Theta_k}}{2} I(\mu_t|\nu_{\tau_{\Theta_k}}). \tag{35}$$

Injecting (35) into (31) and further by Lemma 1, we get

$$\begin{aligned}\frac{d}{dt}H(\mu_t|\nu_{\tau_{\Theta_k}}) &\leq -\frac{3}{2}\tau_{\Theta_k} I(\mu_t|\nu_{\tau_{\Theta_k}}) + C'\left(\eta_{k+1}^2 \Theta_k^{\frac{E_*}{E}} \ln \Theta_k\right) H(\mu_k|\nu_{\tau_{\Theta_k}}) + C'(\eta_{k+1} + \eta_{k+1}^2 \ln \Theta_k) \\ &\leq -\frac{3}{2} C \Theta_k^{-(\frac{E_*}{E} + \varepsilon)} H(\mu_t|\nu_{\tau_{\Theta_k}}) + C'\left(\eta_{k+1}^2 \Theta_k^{\frac{E_*}{E}+\varepsilon} H(\mu_k|\nu_{\tau_{\Theta_k}}) + (\eta_{k+1} + \eta_{k+1}^2 \ln \Theta_k)\right). \end{aligned}$$

Now by a Grönwall argument, we have

$$\begin{aligned}H(\mu_{k+1}|\nu_{\tau_{\Theta_k}}) &\leq e^{-\frac{3}{2}C\eta_{k+1}\Theta_k^{-(\frac{E_*}{E}+\varepsilon)}}\left((1 + C'\eta_{k+1}^3 \Theta_k^{\frac{E_*}{E}+\varepsilon}) H(\mu_k|\nu_{\tau_{\Theta_k}}) + C'(\eta_{k+1}^2 + \eta_{k+1}^3 \ln \Theta_k)\right) \\ &\leq e^{-\frac{5}{4}C\eta_{k+1}\Theta_k^{-(\frac{E_*}{E}+\varepsilon)}} H(\mu_k|\nu_{\tau_{\Theta_k}}) + C'(\eta_{k+1}^2 + \eta_{k+1}^3 \ln \Theta_k) \\ &\leq \left(1 - C\eta_{k+1}\Theta_k^{-(\frac{E_*}{E}+\varepsilon)}\right) H(\mu_k|\nu_{\tau_{\Theta_k}}) + C'(\eta_{k+1}^2 + \eta_{k+1}^3 \ln \Theta_k), \end{aligned} \tag{36}$$

where we use the fact that $\eta_{k+1}\Theta_k^{\frac{E_*}{E}} \to 0$ as $k \to \infty$ in the second inequality.

Now we consider the term (b) in (30). Note that

$$\begin{aligned}H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) - H(\mu_{k+1}|\nu_{\tau_{\Theta_k}}) &= \ln\left(\frac{Z_{\tau_{\Theta_{k+1}}}}{Z_{\tau_{\Theta_k}}}\right) + \left(\frac{1}{\tau_{\Theta_{k+1}}} - \frac{1}{\tau_{\Theta_k}}\right)\mathbb{E}f(x_{k+1}) \\ &\leq C\frac{\eta_{k+1}}{\Theta_k}\mathbb{E}f(x_{k+1}), \end{aligned} \tag{37}$$

since τ_t is decreasing in t, and $\frac{d}{dt}\left(\frac{1}{\tau_t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to \infty$. We claim that for each $\varepsilon > 0$, $\mathbb{E}f(x_{k+1}) \le C \le C\Theta_k^{\varepsilon}$. We argue by contradiction and assume that the sequence $(\mathbb{E}f(x_k),\, k = 0, 1, \ldots)$ is unbounded. Choose $C > 0$ sufficiently large, and let $\mathbb{E}f(x_{k+1})$ be the first term exceeding C. By Assumption 3,

$$f(x_{k+1}) \le f(x_k) - \eta_k |\nabla f(x_k)|^2 + \sqrt{2\tau_{\Theta_k}\eta_k}\nabla f(x_k)\cdot Z_k + \frac{L}{2}|\eta_k \nabla f(x_k) + \sqrt{2\tau_{\Theta_k}\eta_k} Z_k|^2.$$

Further by taking expectation, we get

$$\mathbb{E}f(x_{k+1}) - \mathbb{E}f(x_k) \le -\eta_k\left(1 - \frac{\eta_k L}{2}\right)\mathbb{E}|\nabla f(x_k)|^2 + Ld\tau_{\Theta_k}\eta_k. \tag{38}$$

Thus, $\mathbb{E}f(x_{k+1}) - \mathbb{E}f(x_k) \le Ld\tau_{\Theta_k}\eta_k$ which implies that $\mathbb{E}f(x_k) > C - 1$ for k large enough. Next we prove that $\mathbb{E}|\nabla f(x_j)|^2$ is bounded from below as $j \to \infty$. By Assumption 1, f has quadratic growth at infinity. This implies that for x sufficiently large, say $|x| > R$, we have $|\nabla f(x)|^2 > Af(x)$ for some $A > 0$. Take C' sufficiently large so that $\mathbb{E}f(x_k) > C'$ implies that $\mathbb{E}(f(x_j)1_{\{x_j > R\}}) > C'/2$. Consequently, $\mathbb{E}|\nabla f(x_j)|^2 \ge \mathbb{E}(|\nabla f(x_j)|^2 1_{\{x_j > R\}}) > AC'/2$. Combining with (38), we have $\mathbb{E}f(x_k) > \mathbb{E}f(x_{k+1}) \ge C$. This contradicts the fact that $\mathbb{E}f(x_{k+1})$ is the first term exceeding C. Now by (37), we get

$$H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) - H(\mu_{k+1}|\nu_{\tau_{\Theta_k}}) \le C\eta_{k+1}\Theta_k^{-1+\varepsilon}. \tag{39}$$

Combining (30) with (36), (39) yields (29). □

Step 4: Estimating $H(\mu_k|\nu_{\tau_{\Theta_k}})$. We use Lemma 4 to derive an estimate for $H(\mu_k|\nu_{\tau_{\Theta_k}})$. Under the condition (5), the term $\eta_{k+1}\Theta_k^{-1+\varepsilon}$ dominates η_{k+1}^2, $\eta_{k+1}^3 \ln\Theta_k$ as $k \to \infty$. Thus, the recursion (29) yields

$$H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) \le \left(1 - C\eta_{k+1}\Theta_k^{-(\frac{E_*}{E}+\varepsilon)}\right)H(\mu_k|\nu_{\tau_{\Theta_k}}) + C'\eta_{k+1}\Theta_k^{-1+\varepsilon}.$$

Since $E_*/E < 1$, a similar argument as in Step 4 in Section 3.2 shows that

$$H(\mu_{k+1}|\nu_{\tau_{\Theta_{k+1}}}) - C\Theta_{k+1}^{-(1-\frac{E_*}{E}-2\varepsilon)} \le \left(1 - C'\eta_{k+1}\Theta_k^{-(\frac{E_*}{E}+\varepsilon)}\right)\left(H(\mu_k|\nu_{\tau_{\Theta_k}}) - C\Theta_k^{-(1-\frac{E_*}{E}-2\varepsilon)}\right).$$

Applying the above inequality recursively, we get

$$H(\mu_k|\nu_{\tau_{\Theta_k}}) \le C\Theta_k^{-(1-\frac{E_*}{E}-2\varepsilon)} + \prod_{j=k_0}^{k-1}\left(1 - C'\eta_j\Theta_j^{-(\frac{E_*}{E}+\varepsilon)}\right)\left(H(\mu_k|\nu_{\tau_{\Theta_{k_0}}}) - C\Theta_{k_0}^{-(1-\frac{E_*}{E}-2\varepsilon)}\right). \tag{40}$$

By the sum-integral trick, we have

$$\sum_{j=k_0}^{k}\eta_j\Theta_j^{-(\frac{E_*}{E}+\varepsilon)} \ge \int_{\Theta_{k_0}}^{\Theta_k} z^{-(\frac{E_*}{E}+\varepsilon)}dz,$$

which diverges to ∞ as $E > E_*$ and $\Theta_k \to \infty$ as $k \to \infty$. Combining (40) and the fact that $\prod_j(1 - x_j) \le e^{-\sum_j x_j}$ yields

$$H(\mu_k|\nu_{\tau_{\Theta_k}}) \le C\Theta_k^{-(1-\frac{E_*}{E}-2\varepsilon)}. \tag{41}$$

By injecting (23) and (41) into (22) we obtain (6). The proof is complete.

5. Numerical results

This section presents numerical experiments to corroborate our main result, Theorem 1. We consider two nonconvex functions commonly used in global optimization:

Ackley function: for $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$,

$$f(\boldsymbol{x}) = -a \exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(cx_i)\right) + a + \exp(1), \tag{42}$$

where $a, b, c > 0$ are parameters, and d is the dimension. The Ackley function attains its global minimum at $\boldsymbol{x}^* = (0, \dots, 0)$ with $f(\boldsymbol{x}^*) = 0$. In the sequel, we take $a = 20$, $b = 0.2$, $c = 2\pi$ and $d = 2$ for numerical experiments.

Rastrigin function: for $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$,

$$f(\boldsymbol{x}) = 20 + x_1^2 + x_2^2 - 10\cos(2\pi x_1) - 10\cos(2\pi x_2). \tag{43}$$

The Rastrigin function attains its global minimum at $\boldsymbol{x}^* = (0, 0)$ with $f(\boldsymbol{x}^*) = 0$. See Figure 2 for the landscape of the Ackley and the Rastrigin function in $\mathbb{R}^2$.

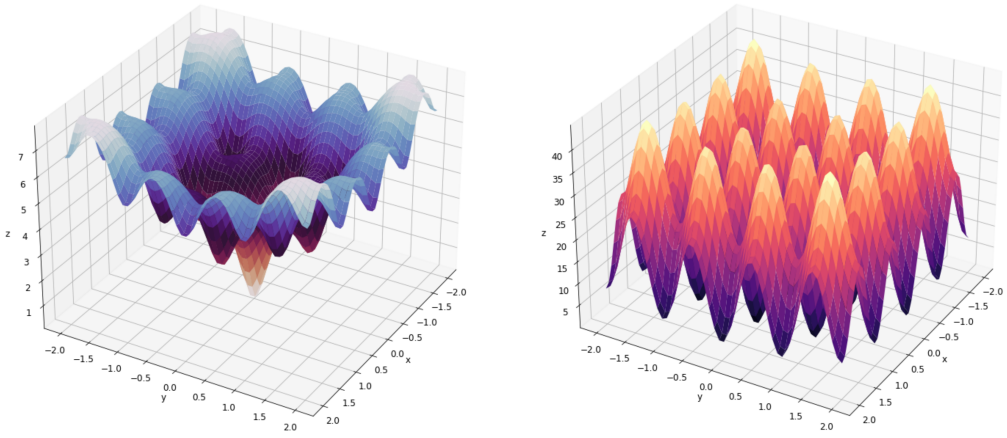


Figure 2. The landscape of the Ackley and the Rastrigin function in $\mathbb{R}^2$. Left: Ackley function with $a = 20$, $b = 0.2$, $c = 2\pi$; Right: Rastrigin function.

Experiments for the Ackley function: It is clear that the Ackley function satisfies Assumptions 1 – 3. Further we choose $\eta_k = k^{-0.7}$, $\Theta_k = \sum_{j\le k}\eta_j$ (so that $\Theta_k \to \infty$ and $\eta_{k+1}\Theta_k \to 0$), and $\tau_{\Theta_k} = \frac{E}{\log(1+\Theta_k)}$ for some range of E. It follows from the deviation bound (6) that

$$\mathbb{P}(f(x_k) \ge \delta) \le \begin{cases} C\Theta_k^{-\frac{1}{2}(1-\frac{E_*}{E})+\epsilon} & \text{for } E_* \le E < E_* + 2\delta, \\ C\Theta_k^{-\frac{\delta}{E}+\epsilon} & \text{for } E \ge E_* + 2\delta. \end{cases}$$

In general, the exact value of E_* is intractable, and to the best of our knowledge, no previous work has considered how to estimate the critical depth E_* of a nonconvex function. With different values of E we expect to observe different patterns of the discrete SA algorithm, and this provides a way to find the numerical value of E_* as we will explain.

We initialize the process with $x_0 = (1.0, 1.0)$, and consider for the range of values $\delta \in \{0.02, 0.03, \dots, 0.50\}$, and $E \in \{0.02, 0.04, 0.06, ..., 2.00\}$. For each pair of ($E$, δ), we run the discrete SA process (1) for 5000 times to generate Monte–Carlo estimates of the tail probabilities for the first 20000 iterations, that is $p^{(k)}_{E,\delta} := \mathbb{P}(f(x_k) \geq \delta)$, $k = 1, 2, ..., 20000$. Denote these estimates

$$P_{E,\delta} = (p^{(1)}_{E,\delta}, p^{(2)}_{E,\delta}, ..., p^{(20000)}_{E,\delta}) \quad \text{and} \quad \Theta = (\Theta_1, \Theta_2, ..., \Theta_{20000}).$$

It is key to note that the discrete SA process may get trapped in a local minimum if the value of E is too small compared to E_*. To illustrate, Figure 3 displays an extract of the Monte–Carlo estimates $P_{E,0.5}$ for different E's.

K	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	13000	14000	15000	16000	17000	18000	19000	20000
E																				
0.02	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418	0.1418
0.04	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508	0.1508
0.06	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368	0.1368
0.08	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324	0.1324
0.10	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300	0.1300
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1.92	0.8110	0.7426	0.7118	0.6800	0.6716	0.6512	0.6402	0.6396	0.6286	0.6146	0.6204	0.6006	0.6092	0.6094	0.5918	0.5942	0.5894	0.5778	0.5886	0.5792
1.94	0.8088	0.7456	0.7274	0.6858	0.6766	0.6556	0.6454	0.6450	0.6394	0.6284	0.6128	0.6096	0.6024	0.6142	0.6086	0.5944	0.5886	0.5948	0.5946	0.5748
1.96	0.8176	0.7548	0.7294	0.6934	0.6932	0.6672	0.6608	0.6498	0.6326	0.6432	0.6264	0.6230	0.6178	0.6140	0.6128	0.6112	0.6092	0.6110	0.5972	0.5806
1.98	0.8228	0.7648	0.7210	0.6982	0.6858	0.6766	0.6604	0.6478	0.6488	0.6462	0.6338	0.6292	0.6178	0.6086	0.6250	0.6072	0.6044	0.6024	0.5930	0.6002
2.00	0.8276	0.7696	0.7348	0.7208	0.6902	0.6606	0.6590	0.6676	0.6486	0.6384	0.6460	0.6290	0.6266	0.6200	0.6262	0.6222	0.6138	0.6012	0.6010	0.6144

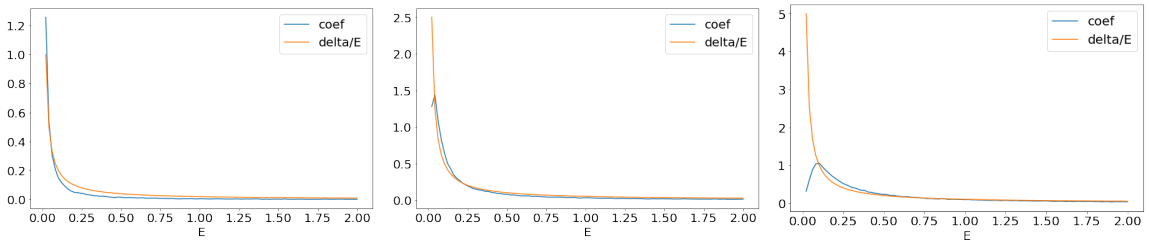
FIGURE 3. Monte Carlo estimates $P_{E,0.5}$ for $E \in \{0.02, 0.04, 0.06, ..., 2.00\}$.

For small E's (e.g. 0.02 – 0.10), the Monte Carlo estimates $p^{(k)}_{E,0.5}$ remains unchanged for all k. The sequence $p^{(k)}_{E,0.5}$, $k = 1, 2, ..., 20000$ is observed to be decreasing from $E = 0.14$ on. This suggests an estimate of E_* which lies between 0.12 and 0.14. Moreover, with the estimates $P_{E,\delta}$ recorded, we take the logarithmic values and run linear regressions of form:

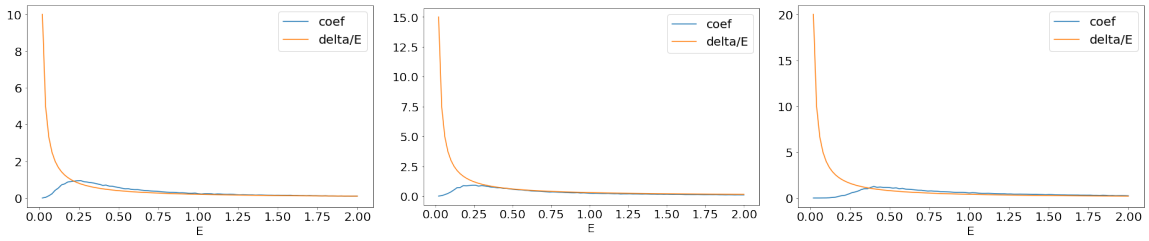
$$\log p^{(k)}_{E,\delta} = \beta_{E,\delta} \log \Theta_k + \gamma_{E,\delta}.$$

The estimated $-\hat{\beta}_{E,\delta}$ then approximates the decay rate of the tail probability $\mathbb{P}(f(x_k) \geq \delta)$ relative to Θ_k. For each pair of (E, δ), we compare $-\hat{\beta}_{E,\delta}$ (dubbed "coef" in the legends) and $\frac{\delta}{E}$. As shown in Figure 4, for all values of δ, as the value of E becomes large, $-\hat{\beta}_{E,\delta}$ fits perfectly with $\frac{\delta}{E}$. We also observe "peaks" in the "coef" curves before they coincide with the respective $\frac{\delta}{E}$ curves. This is due to the fact that for $E_* \leq E < E_* + 2\delta$, the rate $\frac{1}{2}(1 - \frac{E_*}{E})$ increases as E increases. Moreover, the "peaks" occur later when δ is larger since the term $\frac{\delta}{E}$ comes into dominance later.

Experiments for the Rastrigin function: We also choose $\eta_k = k^{-0.7}$, $\Theta_k = \sum_{j \leq k} \eta_j$, and $\tau_{\Theta_k} = \frac{E}{\log(1+\Theta_k)}$ for some range of E. We initialize the process with $x_0 = (1.0, 1.0)$ and consider for the range of values $\delta \in \{0.02, 0.03, \dots, 0.50\}$, and $E \in \{0.05, 0.10, 0.15, ..., 27.00\}$. For each pair of ($E$, δ), we run the discrete SA (1) for 20000 iterations for 5000 times. The results are similar to those for the Ackley function, as displayed in Figure 5. In particular, the



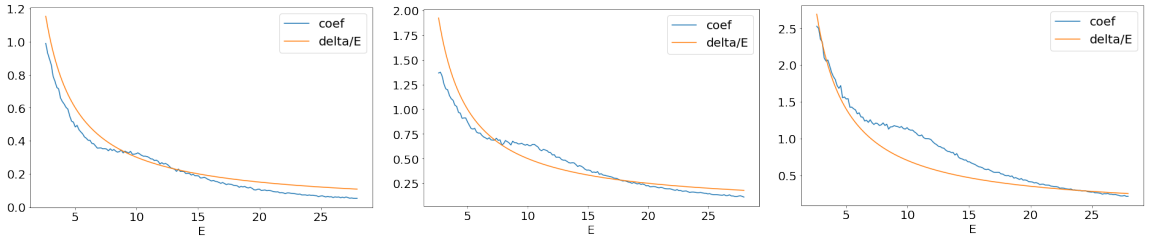
(A) $\delta = 0.02$ (B) $\delta = 0.05$ (C) $\delta = 0.1$



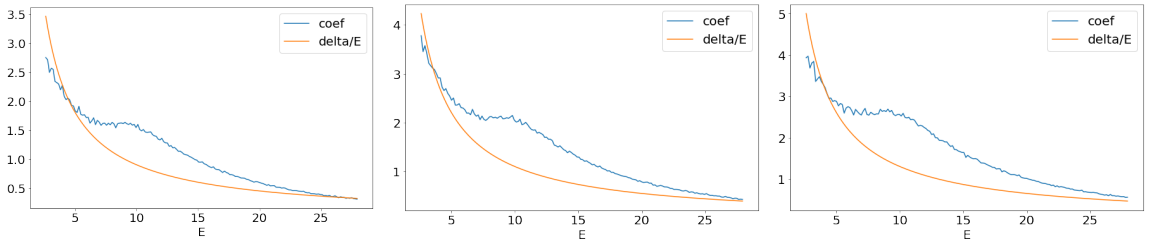
(D) $\delta = 0.2$ (E) $\delta = 0.3$ (F) $\delta = 0.4$

FIGURE 4. Ackley: plots of $\hat{\beta}_{\delta,E}$ and $\frac{\delta}{E}$ against E for different δ's.

estimated E_* lies between 0.05 and 0.10, which is consistent with the fact that the Rastrigin function is *flatter* than the Ackley function.



(A) $\delta = 3$ (B) $\delta = 5$ (C) $\delta = 7$



(D) $\delta = 9$ (E) $\delta = 11$ (F) $\delta = 13$

FIGURE 5. Rastrigin: plots of $\hat{\beta}_{\delta,E}$ and $\frac{\delta}{E}$ against E for different δ's.

6. Conclusion

In this paper, we study the convergence rate of discrete SA processes. The main tool is functional inequalities for the Gibbs measure at low temperatures. We prove that the tail probability exhibits a polynomial decay in time and provide a non-asymptotic deviation bound. The decay rate is given as a function of the model parameters. More importantly, we derive a condition on the step size to ensure the convergence to the global minimum. This condition is useful in tuning the step size as illustrated by numerical experiments.

There are a few directions to extend this work. One is to study the discrete SA with heavy-tailed perturbation under a suitable cooling schedule. Another direction is to study the dependence of the convergence rate in the dimension d. Both problems are challenging but worth exploring.

Acknowledgement: We thank Ruocheng Wu for providing the argument in Step 4 in the proof of Theorem 1, which allows us to remove an unnecessary assumption in a previous version of the paper. We also thank Xuedong He, Georg Menz and Xiaolu Tan for helpful discussions, and Cédric Josz for pointing out the reference (Gelfand and Mitter, 1991). Tang gratefully acknowledges financial support through NSF grants DMS-2113779 and DMS-2206038, and through a start-up grant at Columbia University. Zhou gratefully acknowledges financial supports through a start-up grant at Columbia University and through the Nie Center for Intelligent Asset Management.

References

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.

A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc.*, 6(4): 399–424, 2004.

A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. *J. Eur. Math. Soc.*, 7(1):69–99, 2005.

V. Cerny. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *J. Optim. Theory Appl.*, 45(1):41–51, 1985.

X. Chen, S. S. Du, and X. T. Tong. On stationary-point hitting time and ergodicity of stochastic gradient Langevin dynamics. *J. Mach. Learn. Res.*, 21:Paper No. 68, 41, 2020.

Y. Chen, J. Chen, J. Dong, J. Peng, and Z. Wang. Accelerating nonconvex learning via replica exchange Langevin diffusion. In *International Conference on Learning Representations (ICLR)*, 2019.

T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. Diffusion for global optimization in $\mathbf{R}^n$. *SIAM J. Control Optim.*, 25(3):737–753, 1987.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017.

D. Delahaye, S. Chaimatanan, and M. Mongeau. Simulated annealing: from basics to applications. In *Handbook of metaheuristics*, volume 272 of *Internat. Ser. Oper. Res. Management Sci.*, pages 1–35. Springer, 2019.

J. Dong and X. T. Tong. Replica exchange for non-convex optimization. *J. Mach. Learn. Res.*, 22:Paper No. 173, 59, 2021.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

X. Gao, M. Gürbüzbalaban, and L. Zhu. Global convergence of stochastic gradient Hamiltonian Monte Carlo for nonconvex stochastic optimization: nonasymptotic performance bounds and momentum-based acceleration. *Oper. Res.*, 70(5):2931–2947, 2022.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbf{R}^d$. *SIAM J. Control Optim.*, 29(5):999–1018, 1991.

S. Geman and C.-R. Hwang. Diffusions for global optimization. *SIAM J. Control Optim.*, 24(5):1031–1043, 1986.

U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *J. Roy. Statist. Soc. Ser. B*, 56(4):549–603, 1994.

X. Guo, J. Han, M. Tajrobehkar, and W. Tang. Perturbed gradient descent with occupation time. 2020. arXiv:2005.04507.

Y. Hu, X. Wang, X. Gao, M. Gürbüzbalaban, and L. Zhu. Non-convex optimization via non-reversible stochastic gradient Langevin dynamics. 2020. arXiv:2004.02823.

C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *ICML*, pages 1724–1732, 2017.

S. Kirkpatrick, J. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220 (4598):671–680, 1983.

C. Koulamas, S. Antony, and R. Jaen. A survey of simulated annealing applications to operations research problems. *Omega*, 22(1):41–56, 1994.

Y. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proc. Natl. Acad. Sci. USA*, 116(42):20881–20885, 2019.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942–1992, 2021.

G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 2014.

G. Menz, A. Schlichting, W. Tang, and T. Wu. Ergodicity of the infinite swapping algorithm at low temperature. 2018. arXiv:1811.10174.

L. Miclo. Recuit simulé sur $\mathbb{R}^n$. Étude de l'évolution de l'énergie libre. *Annales de l'Institut Henri Poincaré*, 28(2):235–266, 1992.

R. B. Myerson. *Game theory*. Harvard University Press, 1991.

F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000.

G. Parisi. Correlation functions and computer simulations. *Nuclear Phys. B*, 180(3, FS 2): 378–384, 1981.

M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Probab.*, 8(1):10–44, 1998.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *COLT*, pages 1674–1703, 2017.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

G. Royer. *An initiation to logarithmic Sobolev inequalities*, volume 14 of *SMF/AMS Texts and Monographs*. American Mathematical Society, 2007.

Z. Shun and P. McCullagh. Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B*, 57(4):749–760, 1995.

W. Tang and X. Y. Zhou. Tail probability estimates of continuous-time simulated annealing processes. *Numer. Algebra Control Optim.*, 13(3&4):473–485, 2023.

P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated annealing: theory and applications*, volume 37 of *Mathematics and its Applications*. D. Reidel Publishing Co., 1987.

S. Vempala and A. Wibisono. Rapid convergence of the Unadjusted Langevin Algorithm: isoperimetry suffices. In *NeurIPS*, volume 32, pages 8094–8106, 2019.

Y. Wang and S. Wu. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *J. Mach. Learn. Res.*, 21: Paper No. 199, 103, 2020.

Department of Industrial Engineering and Operations Research, Columbia University.
Email address: wt2319@columbia.edu

Columbia University.
Email address: yuhang.wu@columbia.edu

Department of Industrial Engineering and Operations Research, Columbia University.
Email address: xz2574@columbia.edu