Manuscript submitted to AIMS Journal

doi:10.3934/xx.xxxxxxx

TAIL PROBABILITY ESTIMATES OF CONTINUOUS-TIME SIMULATED ANNEALING PROCESSES

Wenpin Tang

Department of Industrial Engineering and Operations Research
Columbia University, USA

Xun Yu Zhou*

Department of Industrial Engineering and Operations Research
Columbia University, USA

(Communicated by the associate editor name)

Abstract. We study the convergence rate of a continuous-time simulated annealing process $(X_t;\, t \geq 0)$ for approximating the global optimum of a given function f. We prove that the tail probability $\mathbb{P}(f(X_t) > \min f + \delta)$ decays polynomial in time with an appropriately chosen cooling schedule of temperature, and provide an explicit convergence rate through a non-asymptotic bound. Our argument applies recent development of the Eyring-Kramers law on functional inequalities for the Gibbs measure at low temperatures.

1. **Introduction.** Simulated annealing (SA) is an umbrella term for a set of stochastic optimization methods. The goal of SA 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 machine learning; see e.g. [10, 25, 47]. The name is inspired by annealing in metallurgy, which is a process aiming to increase the size of crystals by controlled heating and cooling. The stochastic version of SA was independently proposed by [6] and [24]. The idea is as follows: consider a stochastic process related to f which is subject to thermal noise. When simulating this process, one applies a higher temperature initially to escape from saddle points and local optima, and decreases the temperature slowly over time for the process to converge to the global minimum of f with high probability. This works generally if the cooling is slow enough, and the problem is to find the right stochastic process with the fastest possible cooling schedule that approximates the global optimum.

In this paper, we explore the convergence rate of continuous-time SA with an appropriately chosen cooling schedule of temperature. To be more precise, define the *continuous-time SA process* $(X_t;\, t \geq 0)$ by

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\tau_t}\, dB_t, \quad X_0 \overset{d}{=} \mu_0(dx), \tag{1}$$

2020 *Mathematics Subject Classification.* Primary: 60J60, 39B62; Secondary: 90C30.

Key words and phrases. Simulated annealing, continuous time, convergence rate, Eyring–Kramers law, functional inequalities, overdamped Langevin equation.

*Corresponding author: Xun Yu Zhou.

where $(B_t; t \geq 0)$ is a standard Brownian motion in $\mathbb{R}^d$, τ_t is a given deterministic cooling schedule of temperature, and $\mu_0(dx)$ is some initial distribution. This formulation was first considered by [17, 19]. If $\tau_t \equiv \tau$ is constant in time, the process (1) is the well-known *overdamped Langevin equation* whose stationary distribution is the Gibbs measure $\nu_\tau(dx) \propto \exp(-f(x)/\tau)dx$. Thus, we sometimes call (1) an *SA adapted overdamped Langevin equation*; see Section 3.1 for more background.

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

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

under suitable conditions on the function f and the cooling schedule τ_t. There are two motivations for studying this problem. First, there are a line of works on the interplay between sampling and optimization ([27, 28, 38]). If $\tau_t \equiv \tau$ is constant in time, the overdamped Langevin equation converges to the Gibbs measure ν_τ; and for τ sufficiently small, the Gibbs measure ν_τ approximates the Dirac mass at the global minimum of f. Accordingly, one aims to approximate $\min f$ by $\mathbb{E}f(X_t^\tau)$ where $(X_t^\tau; t \geq 0)$ is the overdamped Langevin process with a small, *fixed* temperature parameter τ. This way one needs to simulate *multiple* (*many*) sample paths to estimate $\mathbb{E}f(X_t^\tau)$. The advantage of SA is that for a suitable choice of *time-dependent* τ_t, the process X_t converges almost surely to $\min f$ as $t \to \infty$. Thus, one only needs to simulate *one* sample path to approximate $\min f$. Second, there are recent works on various noisy gradient-based algorithms [7, 16, 20, 23], aiming to escape saddle points and find 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 [36] to curriculum learning [2]. Compared to the gradient-based methods, SA sets priority to find the global minimum, if at the cost of a longer exploration time.

The main technical 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 3.2). Let us elaborate. It was shown in [9, 17, 18] that the correct order of τ_t for the process (1) to converge to the global minimum of f is $(\ln t)^{-1}$. In fact, there is a phase transition related to the *critical depth* E_* of the function f:

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

(b) If $E \leq \liminf_{t\to\infty} \tau_t \ln t \leq \limsup_{t\to\infty} \tau_t \ln t < \infty$ with $E > E_*$, then

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

Roughly speaking, the critical depth E_* is the highest hill one needs to climb starting from a local minimum to the global minimum. The formal definition of the critical depth E_* will be given in Assumption 2; but see Figure 1 below for an illustration when f is a double-well function. Part (a) above was proved by [21], who also proved part (b) for f on a *compact* Riemannian manifold with a convergence rate via a Poincaré inequality (PI). But their argument for part (b) does not extend to the Euclidean space $\mathbb{R}^d$ which is inherently non-compact. [33] proved part (b) for f on $\mathbb{R}^d$ and characterized the fastest cooling schedule using the Eyring–Kramers law for the log-Sobolev inequality (LSI). See also [14, 34, 49] for similar results under different conditions on f. However, none of these results derived any precise convergence rate for SA in $\mathbb{R}^d$. [29] was the first to derive a convergence rate of $(f(X_t); t \geq 0)$ to the global minimum of f in $\mathbb{R}^d$ *asymptotically* via a large deviation

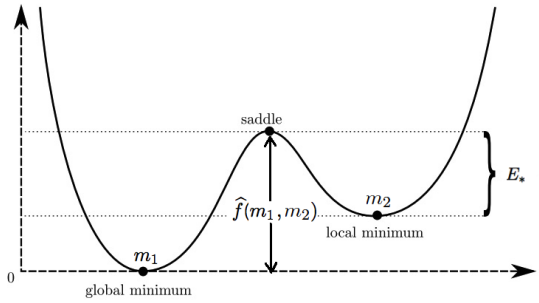

FIGURE 1. Illustration of the critical depth of a double-well function.

principle, i.e. for δ sufficiently small and t sufficiently large. But the bound of $\mathbb{P}(f(X_t) > \min f + \delta)$ for *any* $\delta > 0$ and $t > 0$ has been absent since no estimates of the log-Sobolev inequality for the Gibbs measure at low temperatures were known until the mid-2010s. Taking advantage of some recently developed theory [30, 31], we are able to give a *non-asymptotic* convergence rate of continuous-time SA.

To simplify the notation, we assume henceforth that

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

otherwise we could consider $f - \min f$. Our main result is outlined as follows, whose precise statement will be given in Section 2.

Main Result (Informal)**.** *Under some assumptions on f, and assuming 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 $t \to \infty$, we have for $\delta > 0$, there exists $C > 0$ independent of t such that*

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

This result provides a non-asymptotic bound on the tail probability $\mathbb{P}(f(X_t) > \delta)$ which decays polynomially in time. As mentioned [21] derived a convergence rate for f on a compact Riemannian manifold, with an additional $\log t$ term. This is because they used a weaker PI, while we apply the recently developed Eyring–Kramers formula for LSI which is stronger than the PI. [29] proved a convergence rate for SA in $\mathbb{R}^d$: for $\delta > 0$ sufficiently small and $\tau_t \sim \frac{E}{\ln t}$ with $E > E_*$,

$$\frac{1}{\ln t} \ln \mathbb{P}(f(X_t) > \delta) \to -\frac{\delta}{E} \quad \text{as } t \to \infty. \tag{2}$$

For $\delta > 0$ sufficiently small, we have $\frac{\delta}{E} < \frac{1}{2}(1 - \frac{E_*}{E})$. So our result yields $\mathbb{P}(f(X_t) > \delta) \le C t^{-\frac{\delta}{E}}$, which agrees with the large deviation (2) as $t \to \infty$. On the other hand, our deviation bound holds for all (t, δ, E, E^*). [35] obtained the same rate of convergence for SA adapted *under*damped Langevin equation, and [31] considered an improvement of SA via parallel tempering. However, the (non-asymptotic) convergence rate for SA adapted *over*damped Langevin equation has not appeared in literature to our best knowledge, and here we provide a self-contained treatment that bridges this gap. Moreover, in a separate paper [45], we show how the approach presented in this work can be used to obtain new results in the *discrete-time* setting.

Note that if one uses a *fixed* temperature τ for the overdamped Langevin process $(X_t^\tau; t \geq 0)$, then the tail probability is bounded by

$$\mathbb{P}(f(X_t^\tau) > \delta) \leq \frac{C}{\delta}(\tau + e^{-C't}),$$

for some $C, C' > 0$. The non-vanishing term in τ is inherent – it comes from $\mathbb{E}_{\nu_\tau} f$ in sampling the Gibbs measure $\nu_\tau(dx)$. As a result, the tail probability in this case will *not* converge to 0 over time. Also note that the rate $\min\left(\frac{\delta}{E}, \frac{1}{2}(1 - \frac{E_*}{E})\right)$ for the continuous-time SA is smaller than $\frac{1}{2}$. Empirical results from the discrete setting (see [45]) suggests that this rate is optimal; but it remains open to prove it theoretically. 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. However, the latter may fail if both the dimension d and the inverse temperature $1/\tau$ tend to infinity [42]. As shown in Remark 1 below, we obtain an upper bound for C which is exponential in d. This suggests the convergence rate is exponentially slow as the dimension increases, which concurs the fact that finding the global minimum of a general nonconvex function is NP-hard [22].

Finally, we mention a few approaches in the literature to accelerate or improve SA. [13] considered a cooling schedule depending on both time and state; [37] used the Lévy flight; [35] studied SA adapted to underdamped Langevin equation; [31] applied the replica exchange technique; [15] employed a relaxed stochastic control formulation, originally proposed by [48] for reinforcement learning, to derive a state-dependent temperature control schedule.

The remainder of the paper is organized as follows. Section 2 presents the assumptions and our main results. Section 3 provides background on diffusion processes and functional inequalities. The result for the continuous-time simulated annealing (Theorem 2.1) is proved in Section 4. We conclude with Section **??**.

2. Main results.

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

- The notation $|\cdot|$ is the Euclidean norm of a vector, and $a \cdot b$ is the scalar product of vectors a and b.
- For a function $f : \mathbb{R}^d \to \mathbb{R}$, let ∇f, $\nabla^2 f$ and Δf denote its gradient, Hessian and Laplacian respectively.
- The symbol $a \sim b$ means that $a/b \to 1$ as some problem parameter tends to 0 or ∞. Similarly, the symbol $a = \mathcal{O}(b)$ 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 $(\delta, f, E \ldots)$, and 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, which we take from [30, 31].

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

(i) There exists $C > 0$ such that

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

(ii) There exist $C, C' > 00$ such that

$$\liminf_{|x|\to\infty} \frac{|\nabla f(x)|^2 - \Delta f(x)}{|x|^2} \ge C, \quad \inf_{x\in\mathbb{R}^d} \nabla^2 f(x) \ge -C'.$$

Let us make a few comments on Assumption 1. The condition (i) implies that f is non-degenerate on the set of critical points. 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 which is key to convergence analysis; see [40, Theorem 3.1.21] and Section 3.2. The conditions (i) and (ii) imply that the set of critical points is discrete and finite [30, Remark 1.6]. In particular, it follows that the set of local minimum points $\{m_1, \dots, m_N\}$ is also finite, with N the number of local minimum points of f.

To keep the presentation simple, we make additional assumptions on f, following [31, Assumption 2.5]. 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 $\widehat{f}(m_1, m_2)$ when f is a double-well function with m_1 the global minimum and m_2 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 $m_1, \dots, m_N$ are ordered in the sense that there exists $\delta > 0$ such that

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

(ii) For each $i, j \in \{1, \dots, N\}$, 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 $\{\lambda_1, \dots, \lambda_n\}$ are the eigenvalues of $\nabla^2 f(s_{ij})$, then $\lambda_1 < 0$ and $\lambda_i > 0$ for $i \in \{2, \dots, n\}$. 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 $i \in [N] \setminus \{p\}$,

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

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

The convergence result for the continuous-time SA (1) is stated as follows. The proof will be given in Section 4.

Theorem 2.1. *Let f satisfy Assumptions 1 and 2, and a deterministic function τ_t be 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 the moment condition holds for the initial distribution μ_0: for each $p \ge 1$, there exists $C_p > 0$ such that*

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

Then for each $\delta, \varepsilon > 0$, there exists $C > 0$ independent of t such that

$$\mathbb{P}(f(X_t) > \delta) \le C t^{-\min\left(\frac{\delta}{E}, \frac{1}{2}(1-\frac{E_*}{E})\right)+\varepsilon}. \tag{5}$$

3. **Preliminaries.** In this section, we present a few vocabularies and basic results of diffusion processes and functional inequalities. We also explain how these results are applied in the setting of SA, which will be useful in our convergence analysis.

3.1. **Diffusion processes and SA.** Consider the general diffusion process $(X_t;\, t \ge 0)$ in $\mathbb{R}^d$ of form:

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, \quad X_0 \overset{d}{=} \mu_0(dx), \tag{6}$$

where $(B_t; t \ge 0)$ is a d-dimensional Brownian motion, with the drift $b : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and the covariance matrix $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$. To ensure the well-posedness of the SDE (6), it requires some growth and regularity conditions on b and σ. For instance,

- If b and σ are Lipschitz and have linear growth in x uniformly in t, then (6) has a strong solution which is pathwise unique.
- If b is bounded, and σ is bounded, continuous and strictly elliptic, then (6) has a weak solution which is unique in distribution.

We refer to [43, 39] for background and further developments on the well-posedness of SDEs, and to [8, Chapter 1] for a review of related results.

Another important aspect is the distributional property of $(X_t;\, t \ge 0)$ governed by the SDE (6). Let $\mathcal{L}$ be the infinitesimal generator of the diffusion process X defined by

$$\begin{aligned}\mathcal{L}g(t,x) &:= b(t,x)\cdot\nabla g(x) + \frac{1}{2}\sigma(t,x)\sigma(t,x)^T : \nabla^2 g(x)\\ &:= \sum_{i=1}^d b_i(t,x)\frac{\partial}{\partial x_i}g(x) + \frac{1}{2}\sum_{i,j=1}^d \left(\sigma(t,x)\sigma(t,x)^T\right)_{ij}\frac{\partial^2}{\partial x_i \partial x_j}g(x),\end{aligned} \tag{7}$$

and $\mathcal{L}^*$ be the corresponding adjoint operator given by

$$\begin{aligned}\mathcal{L}^* g(t,x) &:= -\nabla\cdot(b(t,x)g(x)) + \frac{1}{2}\nabla^2 : (\sigma(t,x)\sigma(t,x)^T g(x))\\ &:= -\sum_{i=1}^d \frac{\partial}{\partial x_i}(b_i(t,x)g(x)) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x_j}(\sigma(t,x)\sigma(t,x)^T g(x))_{ij},\end{aligned} \tag{8}$$

where $g : \mathbb{R}^d \to \mathbb{R}$ is a suitably smooth test function, and : denotes the Frobenius inner product which is the component-wise inner product of two matrices. The probability density function $\rho_t(\cdot)$ of the process X at time t then satisfies the Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \mathcal{L}^*\rho_t. \tag{9}$$

Specializing (9) to the SA process (1) with $b(t,x) = -\nabla f(x)$ and $\sigma(t,x) = \sqrt{2\tau_t}\, I_d$, we have that the probability density $\mu_t(\cdot)$ of X governed by the SDE (1) satisfies

$$\frac{\partial \mu_t}{\partial t} = \nabla\cdot(\mu_t \nabla f) + \tau_t \Delta \mu_t. \tag{10}$$

In the time-homogeneous case where $b(t,x) = b(x)$ and $\sigma(t,x) = \sigma(x)$, it can be shown that as $t \to \infty$, $\rho_t(\cdot) \to \rho_\infty(\cdot)$, which is the stationary distribution of $(X_t;\, t \ge 0)$, under further growth conditions on b and σ. It is easily deduced from

(9) that ρ_∞ is characterized by the equation $\mathcal{L}^*\rho_\infty = 0$; see [11, 32] for a general theory on stability of diffusion processes, and [44, Section 2] for a summary with various pointers to the literature. However, for general b and σ, the stationary distribution $\rho_\infty(\cdot)$ does not have a closed-form expression. One good exception is $b(t,x) = -\nabla f(x)$ and $\sigma(t,x) = \sqrt{2\tau}\, I_d$, where X is governed by the overdamped Langevin equation:

$$dX_t = -\nabla f(X_t)dt + \sqrt{2\tau}\, dB_t, \quad X_0 \overset{d}{=} \mu_0(dx). \tag{11}$$

Such a process is time-reversible, and the stationary distribution, under some growth condition on f, is the Gibbs measure

$$\nu_\tau(dx) = \frac{1}{Z_\tau} \exp\left(-\frac{f(x)}{\tau}\right) dx, \tag{12}$$

where $Z_\tau := \int_{\mathbb{R}^d} \exp(-f(x)/\tau)dx$ is the normalizing constant. Much is known about the overdamped Langevin dynamics. For instance, if f is λ-convex (i.e. $\nabla^2 f + \lambda I_d$ is positive definite), the overdamped Langevin process governed by (11) converges exponentially in the Wasserstein metric with rate λ to the Gibbs measure ν_τ [5]. See also [1] for modern techniques to analyze the evolution of the overdamped Langevin equation and generalizations.

Now we turn to the SA process (1). The difference between the overdamped Langevin process (11) and the process (1) is that the temperature τ_t of the latter is decreasing in time. Due to the time dependence, the limiting distribution of SA is unknown. As we will see in Section 4, the idea of analyzing (1) is to approximate it by a process whose stationary distribution is the Gibbs measure with the temperature τ_t. Since τ_t decreases to 0 in the limit, the problem boils down to studying Gibbs measures at low temperatures. In the next section, we recall some results of Gibbs measures at low temperatures, which are originally motivated by applications in molecular dynamics and Bayesian statistics.

3.2. **Functional inequalities and the Eyring–Kramers law.** Here 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 $d\mu/d\nu$ the Radon-Nikodym derivative. Define the relative entropy or KL-divergence $H(\mu|\nu)$ of μ with respect to ν by

$$H(\mu|\nu) := \int \log\left(\frac{d\mu}{d\nu}\right) d\mu = \int \frac{d\mu}{d\nu} \log\left(\frac{d\mu}{d\nu}\right) d\nu, \tag{13}$$

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

$$I(\mu|\nu) := \frac{1}{2} \int \left|\nabla\left(\frac{d\mu}{d\nu}\right)\right|^2 \left(\frac{d\mu}{d\nu}\right)^{-1} d\nu. \tag{14}$$

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

$$H(\mu|\nu) \le \frac{1}{\alpha} I(\mu|\nu). \tag{15}$$

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

The LSI also plays an important role in studying the convergence rate of the overdamped Langevin equation. Recall that ν_τ is the Gibbs measure defined by

(12), and assume that ν_τ satisfies the LSI with constant $\alpha_\tau > 0$. It follows from [41, Theorem 1.7] that by letting $\mu_{\tau,t}$ be the probability distribution of X_t defined by (11), we have

$$H(\mu_{\tau,t}|\nu_\tau) \le e^{-2\tau\alpha_\tau t} H(\mu_{\tau,0}|\nu_\tau). \tag{16}$$

So larger the value of α_τ is, faster the convergence of the overdamped Langevin process in the KL divergence is. 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 [30, Corollary 3.18], who derived a sharp lower bound for α_τ as $\tau \to 0$.

Lemma 3.1. *Let f satisfy Assumptions 1 and 2. Then the Gibbs measure ν_τ defined by (12) satisfies the LSI with constant $\alpha_\tau > 0$ such that*

$$\alpha_\tau \sim C \exp\left(-\frac{E_*}{\tau}\right) \quad \text{as } \tau \to 0, \tag{17}$$

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

The Eyring–Kramers law provides an estimate on the spectral gap of the overdamped Langevin equation (11). It dates back to [12, 26] in the study of metastability in chemical reactions (i.e. mean transition times between local minima and the global one), and is proved rigorously by [3, 4] in terms of the spectral gap. Lemma 3.1 is the LSI version of the Eyring–Kramers law, which is stronger than the spectral gap estimate implied by the Poincaré inequality.

4. **Continuous-time simulated annealing.** In this section, we prove Theorem 2.1 by using the ideas developed in [31, 33, 35]. Let μ_t be the probability measure of X_t defined by (1). The key idea is to compare μ_t with the time-dependent Gibbs measure ν_{τ_t} given by

$$\nu_{\tau_t}(dx) = \frac{1}{Z_{\tau_t}} \exp\left(-\frac{f(x)}{\tau_t}\right) dx, \tag{18}$$

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 2.1 is broken into four steps.

Step 1: Reduce μ_t to ν_{τ_t}. We establish a bound that relates ν_{τ_t} to μ_t. Let $(\widetilde{X}_t;\, t \ge 0)$ be a process (i.e. a collection of random variables parameterized by $t \ge 0$) whose distribution is ν_{τ_t} at time t, defined on the same probability space as $(X_t;\, t \ge 0)$. Fix $\delta > 0$. We have

$$\begin{aligned}
\mathbb{P}(f(X_t) > \delta) &= \mathbb{P}(f(X_t) > \delta, f(\widetilde{X}_t) > \delta) + \mathbb{P}(f(X_t) > \delta, f(\widetilde{X}_t) \le \delta) \\
&\le \mathbb{P}(f(\widetilde{X}_t) > \delta) + \mathbb{P}(X_t \ne \widetilde{X}_t) \\
&\le \mathbb{P}(f(\widetilde{X}_t) > \delta) + \sqrt{2H(\mu_t|\nu_{\tau_t})},
\end{aligned} \tag{19}$$

where the first inequality follows from the fact that $\{f(X_t) > \delta, f(\widetilde{X}_t) \le \delta\} \subset \{X_t \ne \widetilde{X}_t\}$, and and second one follows from Pinsker's inequality [46, Lemma 2.5]. Now 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 $f(\widetilde{X}_t)$. We study the asymptotics of $\mathbb{P}(f(\widetilde{X}_t) > \delta)$ as $t \to \infty$. The following lemma provides a quantitative estimate of how ν_{τ_t}, or equivalently $\widetilde{X}_t$ concentrates on the minimum point of f as $t \to \infty$.

Lemma 4.1. *Let f satisfy Assumption 1 & 2. Assume that $\tau_t \sim \frac{E}{\ln t}$ as $t \to \infty$ with $E > E_*$. For each $\varepsilon \in (0, \delta)$, there exists $C > 0$ independent of t such that*

$$\mathbb{P}(f(\widetilde{X}_t) > \delta) \leq Ct^{-\frac{\delta-\varepsilon}{E}}. \tag{20}$$

Proof. Note that

$$\mathbb{P}(f(\widetilde{X}_t) > \delta) = \frac{\int_{f(x)>\delta} \exp(-f(x)/\tau_t)dx}{\int_{\mathbb{R}^d} \exp(-f(x)/\tau_t)dx}. \tag{21}$$

Under Assumption 1, f has quadratic growth, so at least linear growth at infinity [30, Lemma 3.14]: there exists $C > 0$ such that for R large enough,

$$f(x) \geq \min_{|z|=R} f(z) + C(|x| - R) \quad \text{for } |x| > R.$$

We can also choose R sufficiently large so that $\min_{|z|=R} f(z) > \delta$. Consequently,

$$\begin{aligned}\int_{f(x)>\delta} \exp(-f(x)/\tau_t)dx &= \int_{f(x)>\delta,|x|\leq R} \exp(-f(x)/\tau_t)dx + \int_{f(x)>\delta,|x|>R} \exp(-f(x)/\tau_t)dx \\ &\leq e^{-\frac{\delta}{\tau_t}} \operatorname{Vol}(B_R) + e^{-\frac{\delta}{\tau_t}} \int_{|x|>R} \exp\left(\frac{C(|x|-R)}{\tau_t}\right) dx \\ &= e^{-\frac{\delta}{\tau_t}} (\operatorname{Vol}(B_R) + \mathcal{O}(\tau_t)), \end{aligned} \tag{22}$$

where $\operatorname{Vol}(B_R)$ is the volume of a ball with radius R. Moreover, there exists $r > 0$ such that $f(x) < \varepsilon$ when $|x - m_1| < r$. Thus,

$$\int_{\mathbb{R}^d} \exp(-f(x)/\tau_t)dx > \int_{|x-m_1|<r} \exp(-f(x)/\tau_t)dx \geq e^{-\frac{\varepsilon}{\tau_t}} \operatorname{Vol}(B_r). \tag{23}$$

Injecting (22), (23) into (21) yields (20). □

Remark 1. It is interesting to get a bound for $\mathbb{P}(f(\widetilde{X}_t) > \delta)$ when the dimension d is large. As mentioned in the introduction, the Laplace bound (23) may fail when $d, t \to \infty$ simultaneously. Recall that m_1 is the minimum point of f. By continuity of f, there exists $r > 0$ such that $f(x) < \varepsilon$ when $|x - m_1| < r$. Thus,

$$\begin{aligned}\int_{\mathbb{R}^d} \exp(-f(x)/\tau_t)dx &\geq \int_{|x-m_1|<r} \exp(-f(x)/\tau_t)dx \\ &\geq e^{-\frac{\varepsilon}{\tau_t}} \operatorname{Vol}(B_r). \end{aligned} \tag{24}$$

Further, if $t/e^{\frac{Ed}{CR}} \to \infty$ as $d \to \infty$,

$$\int_{f(x)>\delta} \exp(-f(x)/\tau_t)dx = e^{-\frac{\delta}{\tau_t}} \operatorname{Vol}(B_R)(1 + \mathcal{O}(\tau_t d)). \tag{25}$$

Combining (24) and (25), we get

$$\mathbb{P}(f(\widetilde{X}_t) > \delta) \leq C\gamma^d t^{-\frac{\delta-\varepsilon}{E}}, \tag{26}$$

where $C > 0$ depends on $\delta, \varepsilon, f, E$, and $\gamma = \max(R/r, e^{\frac{\delta-\varepsilon}{CR}})$. Also note that [38] obtained the bound $\mathbb{E}f(\widetilde{X}_t) \leq Cd/\ln t$. By Markov's inequality, we get

$$\mathbb{P}(f(\widetilde{X}_t) > \delta) \leq C\delta^{-1} d\,(\ln t)^{-1}. \tag{27}$$

In comparison with (27), the bound (26) is better in 't' but worse in 'd'. In terms of relaxation time, i.e. letting $\mathbb{P}(f(\widetilde{X}_t) > \delta)$ be of constant order, both estimates show an exponential dependence of t on d. This suggests that SA is exponentially slow as the dimension increases.

Step 3: Differential inequality for $H(\mu_t|\nu_{\tau_t})$. To get an estimate of $H(\mu_t|\nu_{\tau_t})$, we need to consider the time derivative $\frac{d}{dt}H(\mu_t|\nu_{\tau_t})$. The following lemma is a reformulation of [33, Proposition 3]. For ease of reference, we give a simplified proof here. First let us convent some notation. For an absolutely continuous measure $\mu(dx)$, we abuse the notation $\mu(dx) = \mu(x)dx$, i.e. $\mu(x)$ is the density of $\mu(dx)$. So for two such probability measures μ and ν, the Radon-Nikodym derivative $\frac{d\mu}{d\nu}(x)$ is identified with $\frac{\mu(x)}{\nu(x)}$.

Lemma 4.2. *Let τ_t be decreasing in t. We have*

$$\frac{d}{dt}H(\mu_t|\nu_{\tau_t}) \le -2\tau_t I\left(\mu_t|\nu_{\tau_t}\right) + \frac{d}{dt}\left(\frac{1}{\tau_t}\right)\mathbb{E}f(X_t), \tag{28}$$

where $I(\mu_t|\nu_{\tau_t})$ is the Fisher information defined by (14).

Proof. Observe that

$$\begin{aligned}\frac{d}{dt}H(\mu_t|\nu_{\tau_t}) &= \frac{d}{dt}\int \mu_t \ln\left(\frac{\mu_t}{\nu_{\tau_t}}\right)dx \\ &= \underbrace{\int \frac{\partial \mu_t}{\partial t}\ln\left(\frac{\mu_t}{\nu_{\tau_t}}\right)dx}_{(a)} + \underbrace{\int \mu_t \frac{\partial}{\partial t}\left(\ln\left(\frac{\mu_t}{\nu_{\tau_t}}\right)\right)dx}_{(b)}.\end{aligned} \tag{29}$$

We first consider the term (a). Recall that μ_t satisfies the Fokker–Planck equation (10). Together with the fact that $\nabla(\tau_t\nu_{\tau_t}) = -\nu_{\tau_t}\nabla f$, we have

$$\frac{\partial \mu_t}{\partial t} = \nabla\cdot\left(\tau_t\nu_{\tau_t}\nabla\left(\frac{\mu_t}{\nu_{\tau_t}}\right)\right). \tag{30}$$

By injecting (30) into the term (a) and further by integration by parts, we get

$$\begin{aligned}(a) &= \int \nabla\cdot\left(\tau_t\nu_{\tau_t}\nabla\left(\frac{\mu_t}{\nu_{\tau_t}}\right)\right)\ln\left(\frac{\mu_t}{\nu_{\tau_t}}\right)dx \\ &= -\int \tau_t\nu_{\tau_t}\nabla\left(\frac{\mu_t}{\nu_{\tau_t}}\right)\cdot\nabla\ln\left(\frac{\mu_t}{\nu_{\tau_t}}\right)dx \\ &= -\tau_t\int\left|\nabla\left(\frac{\mu_t}{\nu_{\tau_t}}\right)\right|^2\left(\frac{\mu_t}{\nu_{\tau_t}}\right)^{-1}d\nu_{\tau_t} = -2\tau_t I(\mu_t|\nu_{\tau_t}).\end{aligned} \tag{31}$$

Now we consider the term (b). Direct computation leads to

$$\begin{aligned}(b) = \int\left(\frac{\partial\mu_t}{\partial t} - \frac{\mu_t}{\nu_{\tau_t}}\frac{\partial\nu_{\tau_t}}{\partial t}\right)dx &= -\int\frac{\partial}{\partial t}\left(\ln\nu_{\tau_t}\right)d\mu_t \\ &= \int\frac{d}{dt}\left(\ln Z_{\tau_t}\right)d\mu_t + \frac{d}{dt}\left(\frac{1}{\tau_t}\right)\mathbb{E}f(X_t) \\ &\le \frac{d}{dt}\left(\frac{1}{\tau_t}\right)\mathbb{E}f(X_t),\end{aligned} \tag{32}$$

where we use the facts that $\int\mu_t dx = 1$ in the second equality and that τ_t is decreasing in t so $\ln Z_{\tau_t}$ is decreasing in t in the last inequality. Combining (29) with (31) and (32) yields (28). □

Step 4: Estimating $H(\mu_t|\nu_{\tau_t})$ via the Eyring–Kramers law. Note that there are two terms on the right hand side of (28). We start with an estimate of the second term.

Lemma 4.3. *Let f satisfy Assumption 1, and assume that the condition (4) for μ_0 holds. Then, for each $\varepsilon > 0$, there exists $C > 0$ independent of t such that*

$$\mathbb{E}f(X_t) \le C(1+t)^{\varepsilon}. \tag{33}$$

Proof. It is easy to see that Assumption 1 implies Assumption H_1 in [33]. Together with the moment condition (4), the proof follows the line of reasoning in [33, Lemma 2]. □

Now we apply the Eyring–Kramers law, combining with a Grönwall-type argument to bound $H(\mu_t|\nu_{\tau_t})$ for large t.

Lemma 4.4. *Under the assumptions of Theorem 2.1, for each $\varepsilon > 0$, there exists $C > 0$ independent of t such that*

$$H(\mu_t|\nu_{\tau_t}) \le Ct^{-\left(1-\frac{E_*}{E}-\varepsilon\right)}. \tag{34}$$

Proof. Using Lemma 3.1 and the bound (28), we have

$$\frac{d}{dt}H(\mu_t|\nu_{\tau_t}) \le -2\tau_t\alpha_t H(\mu_t|\nu_{\tau_t}) + \frac{C}{t}\,\mathbb{E}f(X_t), \tag{35}$$

where α_t is the LSI constant for the Gibbs measure ν_{τ_t}. By the Eyring–Kramers formula (17), for each $\varepsilon > 0$, there exist $C > 0$ and $t_0 > 0$,

$$2\tau_t\alpha_t \ge Ct^{-\left(\frac{E_*}{E}-\varepsilon\right)} \quad \text{for } t \ge t_0. \tag{36}$$

Combining (35) with (33), (36), we get

$$\frac{d}{dt}H(\mu_t|\nu_{\tau_t}) \le -Ct^{-\left(\frac{E_*}{E}-\varepsilon\right)}H(\mu_t|\nu_{\tau_t}) + C't^{-1+\varepsilon}. \tag{37}$$

Fix $\varepsilon \in (0, \frac{1}{2} - \frac{E_*}{2E})$, let

$$Q(t) := H(\mu_t|\nu_{\tau_t}) - \frac{2C'}{C}t^{-1+\frac{E_*}{E}+2\varepsilon}.$$

Then for t_0 sufficiently large and $t \ge t_0$, we have $\frac{d}{dt}Q(t) \le -Ct^{-\frac{E_*}{E}+\varepsilon}Q(t)$ by (37). This implies that $Q(t) \le Q(t_0)e^{-C\int_{t_0}^t s^{-\frac{E_*}{E}+\varepsilon}ds}$. Thus,

$$H(\mu_t|\nu_{\tau_t}) \le \frac{2C'}{C}t^{-1+\frac{E_*}{E}+2\varepsilon} + H(\mu_{t_0}|\nu_{t_0})e^{-\frac{C}{\kappa}(t^{\kappa}-t_0^{\kappa})}, \tag{38}$$

where $\kappa := 1 - \frac{E_*}{E} - \varepsilon > 0$. Note that the first term on the right hand side of (38) dominates, and the conclusion follows. □

Finally, by injecting (20), (34) into (19), we get the desired estimate (5).

5. **Conclusion.** In this paper, we study the convergence rate of SA in continuous time. The main tool is functional inequalities for the Gibbs measures 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.

There are a few directions to extend this work. For instance, one can study the convergence rate of SA for Lévy flight with a suitable cooling schedule. Another problem is to study the dependence of the convergence rate in the dimension d. This requires a deep understanding of the Eyring–Kramers law in high dimensions, and is related to the Laplace approximation of high dimensional integrals. Both problems are worth exploring, if challenging.

Acknowledgments. We thank Georg Menz for helpful discussions, and thank two anonymous referees for their constructive comments that have led to an improved version of the paper. Tang gratefully acknowledges financial support through an NSF grant DMS-2113779 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

[1] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 2014.
[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.
[3] 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.
[4] 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.
[5] J. A. Carrillo, R. J. McCann, and C. Villani. Contractions in the 2-Wasserstein length space and thermalization of granular media. *Arch. Ration. Mech. Anal.*, 179(2):217–263, 2006.
[6] V. Cerny. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *J. Optim. Theory Appl.*, 45(1):41–51, 1985.
[7] 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.
[8] A. S. Cherny and H.-J. Engelbert. *Singular stochastic differential equations*, volume 1858 of *Lecture Notes in Mathematics*. Springer-Verlag, 2005.
[9] 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.
[10] 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.
[11] S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1986.
[12] H. Eyring. The activated complex in chemical reactions. *J. Chem. Phys.*, 3(2):107–115, 1935.
[13] H. Fang, M. Qian, and G. Gong. An improved annealing method and its large-time behavior. *Stochastic Process. Appl.*, 71(1):55–74, 1997.
[14] N. Fournier and C. Tardif. On the simulated annealing in $\mathbb{R}^d$. 2020. arXiv:2003.06360.
[15] X. Gao, Z. Q. Xu, and X. Y. Zhou. State-dependent temperature control for Langevin diffusions. 2020. arXiv:2005.04507.
[16] 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.
[17] S. Geman and C.-R. Hwang. Diffusions for global optimization. *SIAM J. Control Optim.*, 24 (5):1031–1043, 1986.
[18] B. Gidas. Global optimization via the Langevin equation. In *24th IEEE Conference on Decision and Control*, pages 774–778. IEEE, 1985.
[19] U. Grenander. *Tutorial in pattern theory*. Division of Applied Mathematics. Brown University, 1983.
[20] X. Guo, J. Han, M. Tajrobehkar, and W. Tang. Perturbed gradient descent with occupation time. 2020. arXiv:2005.04507.
[21] R. A. Holley, S. Kusuoka, and D. W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.*, 83(2):333–347, 1989.
[22] P. Jain and P. Kar. Non-convex optimization for machine learning. *Found. Trends Mach. Learn.*, 10:142–336, 2017.
[23] C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *ICML*, pages 1724–1732, 2017.
[24] S. Kirkpatrick, J. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220 (4598):671–680, 1983.

[25] C. Koulamas, S. Antony, and R. Jaen. A survey of simulated annealing applications to operations research problems. *Omega*, 22(1):41–56, 1994.
[26] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.
[27] 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.
[28] 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.
[29] D. Márquez. Convergence rates for annealing diffusion processes. *Ann. Appl. Probab.*, 7(4): 1118–1139, 1997.
[30] G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 2014.
[31] G. Menz, A. Schlichting, W. Tang, and T. Wu. Ergodicity of the infinite swapping algorithm at low temperature. 2018. arXiv:1811.10174.
[32] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.*, 25(3):518–548, 1993.
[33] 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.
[34] L. Miclo. Une étude des algorithmes de recuit simulé sous-admissibles. *Ann. Fac. Sci. Toulouse Math. (6)*, 4(4):819–877, 1995.
[35] P. Monmarché. Hypocoercivity in metastable settings and kinetic simulated annealing. *Probability Theory and Related Fields*, pages 1–34, 2018.
[36] R. B. Myerson. *Game theory*. Harvard University Press, 1991.
[37] I. Pavlyukevich. Lévy flights, non-local search and simulated annealing. *J. Comput. Phys.*, 226(2):1830–1844, 2007.
[38] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *COLT*, pages 1674–1703, 2017.
[39] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2, Itô Calculus*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1987.
[40] G. Royer. *An initiation to logarithmic Sobolev inequalities*, volume 14 of *SMF/AMS Texts and Monographs*. American Mathematical Society, 2007.
[41] A. Schlichting. *The Eyring-Kramers formula for Poincaré and logarithmic Sobolev inequalities*. PhD thesis, Universität Leipzig, 2012. Available at http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-97965.
[42] Z. Shun and P. McCullagh. Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B*, 57(4):749–760, 1995.
[43] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*, volume 233 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 1979.
[44] W. Tang. Exponential ergodicity and convergence for generalized reflected Brownian motion. *Queueing Syst.*, 92(1-2):83–101, 2019.
[45] W. Tang, Y. Wu, and X. Y. Zhou. Discrete simulated annealing: a convergence analysis via the Eyring–Kramers law. 2021. arXiv:2102.02339.
[46] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, 2009.
[47] 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.
[48] H. Wang, T. Zariphopoulou, and X. Y. Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21:1–34, 2020.
[49] P. A. Zitt. Annealing diffusions in a potential function with a slow growth. *Stochastic Process. Appl.*, 118(1):76–119, 2008.

Received xxxx 20xx; revised xxxx 20xx; early access xxxx 20xx.

E-mail address: wt2319@columbia.edu
E-mail address: xz2574@columbia.edu