Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Nonconvex Stochastic Optimization: Nonasymptotic Performance Bounds and Momentum-Based Acceleration

Xuefeng Gao, Mert Gürbüzbalaban, Lingjiong Zhu

To cite this article:
Xuefeng Gao, Mert Gürbüzbalaban, Lingjiong Zhu (2021) Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Nonconvex Stochastic Optimization: Nonasymptotic Performance Bounds and Momentum-Based Acceleration. Operations Research

Published online in Articles in Advance 22 Oct 2021

. https://doi.org/10.1287/opre.2021.2162

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–17
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Nonconvex Stochastic Optimization: Nonasymptotic Performance Bounds and Momentum-Based Acceleration

Xuefeng Gao,[a] Mert Gürbüzbalaban,[b] Lingjiong Zhu[c]

[a] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; [b] Department of Management Science and Information Systems, Rutgers Business School, Piscataway, New Jersey 08854; [c] Department of Mathematics, Florida State University, Tallahassee, Florida 32306

Contact: xfgao@se.cuhk.edu.hk, https://orcid.org/0000-0003-2424-8257 (XG); mg1366@rutgers.edu, https://orcid.org/0000-0002-0575-2450 (MG); zhu@math.fsu.edu (LZ)

Received: August 21, 2019
Revised: November 20, 2020
Accepted: April 6, 2021
Published Online in Articles in Advance: October 22, 2021

Subject Classifications: optimization, programming: stochastic; statistics: sampling; probability: diffusion; probability: Markov processes; programming: nonlinear
Area of Review: Data Science

https://doi.org/10.1287/opre.2021.2162

Abstract. Stochastic gradient Hamiltonian Monte Carlo (SGHMC) is a variant of stochastic gradients with momentum where a controlled and properly scaled Gaussian noise is added to the stochastic gradients to steer the iterates toward a global minimum. Many works report its empirical success in practice for solving stochastic nonconvex optimization problems; in particular, it has been observed to outperform overdamped Langevin Monte Carlo–based methods, such as stochastic gradient Langevin dynamics (SGLD), in many applications. Although the asymptotic global convergence properties of SGHMC are well known, its finite-time performance is not well understood. In this work, we study two variants of SGHMC based on two alternative discretizations of the underdamped Langevin diffusion. We provide finite-time performance bounds for the global convergence of both SGHMC variants for solving stochastic nonconvex optimization problems with explicit constants. Our results lead to nonasymptotic guarantees for both population and empirical risk minimization problems. For a fixed target accuracy level on a class of nonconvex problems, we obtain complexity bounds for SGHMC that can be tighter than those available for SGLD.

Funding: M. Gürbüzbalaban's research is supported in part by the grants Office of Naval Research [Award Number N00014-21-1-2244] and the National Science Foundation (NSF) [Grants CCF-1814888, NSF DMS-2053485, and NSF DMS-1723085]. X. Gao acknowledges support from Hong Kong RGC [Grants 14201117 and 14201520]. L. Zhu is grateful to the support from a Simons Foundation Collaboration Grant and the National Science Foundation [Grant NSF DMS-2053454].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2021.2162.

Keywords: Langevin dynamics • stochastic gradient methods • momentum-based acceleration • nonconvex optimization • empirical risk minimization • Gibbs sampling

1. Introduction

We consider the stochastic nonconvex optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) := \mathbb{E}_{Z\sim\mathcal{D}}[f(x,Z)], \tag{1}$$

where Z is a random variable whose probability distribution $\mathcal{D}$ is unknown, supported on some unknown set $\mathcal{Z}$, and the objective F is the expectation of a random function $f:\mathbb{R}^d\times\mathcal{Z}\to\mathbb{R}$, where the functions $x\mapsto f(x,z)$ are continuous and nonconvex. Having access to independent and identically distributed (i.i.d.) samples $\mathbf{Z}=(Z_1,Z_2,\ldots,Z_n)$, where each Z_i is a random variable distributed with the population distribution $\mathcal{D}$, the goal is to compute an approximate minimizer $\hat{x}$ (possibly with a randomized algorithm) of the *population risk*; that is, we want to compute $\hat{x}$ such that $\mathbb{E}F(\hat{x})-F^*\le\hat{\varepsilon}$ for a given target accuracy $\hat{\varepsilon}>0$, where $F^*=\min_{x\in\mathbb{R}^d}F(x)$ is the minimum value and the expectation is taken with respect to both $\mathbf{Z}$ and the randomness encountered (if any) during the iterations of the algorithm to compute $\hat{x}$. This formulation arises frequently in several contexts, including machine learning. A prominent example is deep learning in which x denotes the set of trainable weights for a deep learning model and $f(x,z_i)$ is the penalty (loss) of prediction using weight x with the individual sample value $Z_i=z_i\in\mathcal{Z}$.

Because the population distribution $\mathcal{D}$ is unknown, a common popular approach is to consider the *empirical risk minimization* (ERM) problem

$$\min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,z_i), \tag{2}$$

based on the data set $\mathbf{z} := (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$ as a proxy to Problem (1) and minimize the *empirical risk*

$$\mathbb{E}\, F_{\mathbf{z}}(x) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \tag{3}$$

instead, for which the expectation is taken with respect to any randomness encountered during the algorithm to generate x.[1] Many algorithms have been proposed to solve Problem (1) and its finite-sum version (2). Among these, gradient descent, stochastic gradient, and their variance-reduced or momentum-based variants come with guarantees for finding a local minimizer or a stationary point for nonconvex problems. In some applications, convergence to a local minimum can be satisfactory (Du et al. 2018, Ge et al. 2018). However, in general, methods with global convergence guarantees are also desirable and preferable in many settings (Hazan et al. 2016, Şimşekli et al. 2018).

It is well known that sampling from a distribution that concentrates around a global minimizer of F is a similar goal to computing an approximate global minimizer of F. For example, such connections arise in the study of simulated annealing algorithms in optimization that admit several asymptotic convergence guarantees (see, e.g., Kirkpatrick et al. 1983, Gidas 1985, Hajek 1985, Gelfand and Mitter 1991, Bertsimas and Tsitsiklis 1993, Borkar and Mitter 1999, Belloni et al. 2015). Recent studies make such connections between the fields of statistics and optimization stronger, justifying and popularizing the use of Langevin Monte Carlo–based methods in stochastic nonconvex optimization and large-scale data analysis further (see, e.g., Welling and Teh 2011; Chen et al. 2016; Şimşekli et al. 2016, 2018; Chaudhari et al. 2017; Dalalyan 2017; Raginsky et al. 2017; Wibisono 2018).

Stochastic gradient algorithms based on Langevin Monte Carlo are popular variants of stochastic gradients that admit asymptotic global convergence guarantees with which a properly scaled Gaussian noise is added to the gradient estimate. Two popular Langevin-based algorithms that have demonstrated empirical success are stochastic gradient Langevin dynamics (SGLD) (Welling and Teh 2011, Chen et al. 2015) and stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Duane et al. 1987; Neal 2010; Chen et al. 2014, 2015) and their variants to improve their efficiency and accuracy (Ahn et al. 2012, Patterson and Teh 2013, Ding et al. 2014, Ma et al. 2015, Wibisono 2018). In particular, SGLD can be viewed as the analogue of stochastic gradients in the Markov Chain Monte Carlo (MCMC) literature, whereas SGHMC is the analogue of stochastic gradients with momentum (see, e.g., Chen et al. 2014). SGLD iterations consist of

$$X_{k+1} = X_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k\,,$$

where $\eta > 0$ is the step-size parameter, $\beta > 0$ is the inverse temperature, g_k is a conditionally unbiased estimate of the gradient of $F_{\mathbf{z}}$, and $\xi_k \in \mathbb{R}^d$ is a sequence of i.i.d. centered Gaussian random vectors with unit covariance matrix. When the gradient variance is zero, SGLD dynamics correspond to an (explicit) Euler discretization of the first order (aka overdamped) Langevin stochastic differential equation (SDE)

$$dX(t) = -\nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\beta^{-1}}dB(t)\,, \quad t \geq 0\,, \tag{4}$$

where $\{B(t) : t \geq 0\}$ is the standard Brownian motion in $\mathbb{R}^d$. The process X admits a unique stationary distribution $\pi_{\mathbf{z}}(dx) \propto \exp(-\beta F_{\mathbf{z}}(x))dx$, also known as the *Gibbs measure*, under some assumptions on $F_{\mathbf{z}}$ (see, e.g., Chiang et al. 1987, Holley et al. 1989). For β chosen properly (large enough), it is easy to see that this distribution concentrates around approximate global minimizers of $F_{\mathbf{z}}$. Recently, Dalalyan (2017) established novel theoretical guarantees for the convergence of the overdamped Langevin MCMC and the SGLD algorithm for sampling from a smooth and log-concave density, and these results have direct implications to stochastic convex optimization; see also Dalalyan and Karagulyan (2019). In a seminal work, Raginsky et al. (2017) show that SGLD iterates track the overdamped Langevin SDE closely and obtained finite-time performance bounds for SGLD. Their results show that SGLD converges to ε-approximate global minimizers after $\mathcal{O}(\mathrm{poly}(\frac{1}{\lambda_*}, \beta, d, \frac{1}{\varepsilon}))$ iterations in which λ_* is the uniform spectral gap that controls the convergence rate of the overdamped Langevin diffusion, which is, in general, exponentially small in both β and the dimension d (Raginsky et al. 2017, Tzen et al. 2018). A related result of Zhang et al. (2017a) shows that a modified version of the SGLD algorithm finds an ε-approximate local minimum after polynomial time (with respect to all parameters). Recently, Xu et al. (2018) improved the ε dependency of the upper bounds of Raginsky et al. (2017) further in the mini-batch setting and obtained several guarantees for the gradient Langevin dynamics and variance-reduced SGLD algorithms.

On the other hand, the SGHMC algorithm is based on the underdamped (aka second order or kinetic) Langevin diffusion

$$dV(t) = -\gamma V(t)dt - \nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\gamma\beta^{-1}}dB(t), \tag{5}$$

$$dX(t) = V(t)dt, \tag{6}$$

where $\gamma > 0$ is the friction coefficient, $X(t), V(t) \in \mathbb{R}^d$ models the position and the momentum of a particle moving in a field of force (described by the gradient of $F_{\mathbf{z}}$) plus a random (thermal) force described by Brownian noise, first derived by Kramers (1940). It is known that, under some assumptions on $F_{\mathbf{z}}$, the

Markov process $(X(t),V(t))_{t\geq 0}$ is ergodic and admits a unique stationary distribution

$$\pi_{\mathbf{z}}(dx,dv)=\frac{1}{\Gamma_{\mathbf{z}}}\exp\left(-\beta\left(\frac{1}{2}\|v\|^2+F_{\mathbf{z}}(x)\right)\right)dxdv, \tag{7}$$

(see, e.g., Hérau and Nier 2004, Pavliotis 2014) in which $\Gamma_{\mathbf{z}}$ is the normalizing constant:

$$\begin{aligned}\Gamma_{\mathbf{z}}&=\int_{\mathbb{R}^d\times\mathbb{R}^d}\exp\left(-\beta\left(\frac{1}{2}\|v\|^2+F_{\mathbf{z}}(x)\right)\right)dxdv\\&=\left(\frac{2\pi}{\beta}\right)^{d/2}\int_{\mathbb{R}^d}e^{-\beta F_{\mathbf{z}}(x)}dx.\end{aligned}$$

Hence, the x-marginal distribution of stationary distribution $\pi_{\mathbf{z}}(dx,dv)$ is exactly the invariant distribution of the overdamped Langevin diffusion.[2] SGHMC dynamics correspond to the discretization of the underdamped Langevin SDE in which the gradients are replaced with their unbiased estimates. Although various discretizations of the underdamped Langevin SDE have also been considered and studied (Chen et al. 2015, Leimkuhler et al. 2015), the following first order Euler scheme is the simplest approach that is easy to implement and a common scheme among practitioners (Chen et al. 2015, 2016; Teh et al. 2016):

$$V_{k+1}=V_k-\eta[\gamma V_k+g(X_k,U_{\mathbf{z},k})]+\sqrt{2\gamma\beta^{-1}\eta}\xi_k, \tag{8}$$

$$X_{k+1}=X_k+\eta V_k, \tag{9}$$

where $(\xi_k)_{k=0}^{\infty}$ is a sequence of i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$ and $\{U_{\mathbf{z},k}:k=0,1,\dots\}$ is a sequence of i.i.d. random elements such that

$$\mathbb{E}g(x,U_{\mathbf{z},k})=\nabla F_{\mathbf{z}}(x)\quad\text{for any}\quad x\in\mathbb{R}^d.$$

In this paper, we focus on the unadjusted dynamics (without the Metropolis–Hastings type of correction) that works well in many applications (Chen et al. 2014, 2015) as Metropolis–Hastings correction is typically computationally expensive for applications in machine learning and large-scale optimization when the size of the data set n is large and low-to-medium accuracy is enough in practice (see, e.g., Welling and Teh 2011, Chen et al. 2014).

There is also an alternative discretization to (8) and (9), recently proposed by Cheng et al. (2018a) that leads to state-of-the-art estimates in the special case that improves upon the Euler discretization when the objective is strongly convex (Cheng et al. 2018a). To introduce this alternative discretization by Cheng et al. (2018a), we first define a sequence of functions ψ_k by $\psi_0(t)=e^{-\gamma t}$ and $\psi_{k+1}(t)=\int_0^t\psi_k(s)ds$, $k\geq 0$. The iterates $(\hat{X}_k,\hat{V}_k)$ are then defined by the following recursion:

$$\hat{V}_{k+1}=\psi_0(\eta)\hat{V}_k-\psi_1(\eta)g(\hat{X}_k,U_{\mathbf{z},k})+\sqrt{2\gamma\beta^{-1}}\xi_{k+1}, \tag{10}$$

$$\hat{X}_{k+1}=\hat{X}_k+\psi_1(\eta)\hat{V}_k-\psi_2(\eta)g(\hat{X}_k,U_{\mathbf{z},k})+\sqrt{2\gamma\beta^{-1}}\xi'_{k+1}, \tag{11}$$

where (ξ_{k+1},ξ'_{k+1}) is a $2d$-dimensional centered Gaussian vector so that the (ξ_j,ξ'_j) s are i.i.d. and independent of the initial condition, and for any fixed j, the random vectors $((\xi_j)_1,(\xi'_j)_1)$, $((\xi_j)_2,(\xi'_j)_2)$, … $((\xi_j)_d,(\xi'_j)_d)$ are i.i.d. with the covariance matrix

$$C(\eta)=\int_0^{\eta}[\psi_0(t),\psi_1(t)]^T[\psi_0(t),\psi_1(t)]dt. \tag{12}$$

In the rest of the paper, we refer to the Euler discretization (8) and (9) as SGHMC1 and the alternative discretization (10) and (11) as SGHMC2.

Recently, Eberle et al. (2019) showed that the underdamped SDE converges to its stationary distribution faster than that of the best known convergence rate of overdamped SDE in the two-Wasserstein metric under some assumptions, where $F_{\mathbf{z}}$ can be nonconvex. Their result is for the continuous-time underdamped dynamics. This raises the natural question of whether the discretized underdamped dynamics (SGHMC) can lead to better guarantees than the SGLD method for solving stochastic nonconvex optimization problems. Indeed, experimental results show that SGHMC can outperform SGLD dynamics in many applications (see, e.g., Chen et al. 2014, 2015; Eberle et al. 2019). Although asymptotic convergence guarantees for SGHMC exist (see, e.g., Mattingly et al. 2002, Section 3; Chen et al. 2014; Leimkuhler et al. 2015), there is a lack of finite-time explicit performance bounds for solving nonconvex stochastic optimization problems with SGHMC in the literature including risk minimization problems.

1.1. Contributions

Our main contributions can be summarized as follows:

- We provide, for the first time to our knowledge, nonasymptotic provable guarantees for SGHMC to find approximate minimizers of both empirical and population risks with explicit constants. We establish the results under some regularity and growth assumptions for the component functions $f(x, z)$ and the noise in the gradients, but we do not assume f is strongly convex in any region.
- We show that, for a class of nonconvex problems, SGHMC2 can improve upon the (vanilla) SGLD algorithm in terms of the *gradient complexity*, that is the total number of stochastic gradients required to achieve a global minimum. Here, "improvement" means the best available bounds for SGHMC2, which we prove in our paper, are better than the best available bounds for SGLD for some class of problems; see Section 5 for details. As a consequence, our analysis gives further theoretical justification to the success of momentum-based methods for solving nonconvex machine learning

problems, empirically observed in practice (see, e.g., Sutskever et al. 2013).

• We illustrate the applications of our theoretical results using two examples including binary linear classification and robust ridge regression.

• On the technical side, we adapt the proof techniques of Raginsky et al. (2017) developed for the overdamped dynamics to the underdamped dynamics and combine them with the analysis of Eberle et al. (2019), which quantifies the convergence rate of the underdamped Langevin SDE to its equilibrium. The main new technical results we derive in this paper, relative to these studies, include controlling the discretization errors between SGHMC and the continuous-time underdamped Langevin SDE and bounding the moments of underdamped dynamics.

1.2. Related Work and Comparison with Existing Literature

In a recent work, Şimşekli et al. (2018) obtains a finite-time performance bound for the ergodic average of the SGHMC iterates in the presence of delays in gradient computations. Their analysis highlights the dependency of the optimization error on the delay in the gradient computations and the step size explicitly; however, it hides some implicit constants that can be exponential in both β and d in the worst case. A comparison with the SGLD algorithm is also not given. On the contrary, in our paper, we make all the constants explicit. This allows us to make gradient complexity comparisons with respect to overdamped MCMC approaches, such as SGLD.

Cheng et al. (2018b) consider the problem of sampling from a target distribution $p(x) \propto \exp(-F(x))$, where $F : \mathbb{R}^d \to \mathbb{R}$ is L-smooth everywhere and m-strongly convex outside a ball of finite radius $\mathcal{R}$. They prove upper bounds for the time required to sample from a distribution that is within ε of the target distribution with respect to the one-Wasserstein distance for both underdamped and overdamped methods that scale polynomially in ε and d. They also show that an underdamped MCMC has a better dependency with respect to ε and d by a square root factor. Compared with this paper, in our analysis, we consider a larger class of nonconvex functions $F(x)$ that satisfy the dissipativity condition, a weaker condition that does not require strong convexity outside a region. Under our assumptions, the best known bounds are such that the distance to the invariant distribution scales exponentially with dimension d in the worst case but not polynomially in d (see, e.g., Raginsky et al. 2017, Xu et al. 2018). When F is globally strongly convex (or, equivalently, when the target distribution $p(x) \propto \exp(-F(x))$ is strongly log-concave), there is also a growing interesting literature that establishes performance bounds for both overdamped MCMC (see, e.g., Dalalyan 2017) and underdamped MCMC methods (see, e.g., Mangoubi and Smith 2017, Cheng et al. 2018b). In this particular setting, the fact that underdamped Langevin MCMC (also known as Hamiltonian MCMC) can improve upon the best available bounds for overdamped Langevin MCMC algorithms is also proven (Mangoubi and Smith 2017, Chatterji et al. 2018, Cheng et al. 2018b, Dalalyan and Riou-Durand 2020). Similar results have also been established when $F(x)$ is convex but not strongly convex (Dalalyan et al. 2019). Compared with these papers in which $F(x)$ is convex, our assumptions are weaker as we allow $F(x)$ to be nonconvex as long as it is dissipative.

A related paper, Xu et al. (2018) applies variance reduction techniques to overdamped MCMC to improve performance when the empirical risk can be nonconvex, satisfying the same dissipativity assumption considered in our paper. However, these results do not give guarantees for the risk minimization Problem (1). Furthermore, such variance-reduction techniques require objectives in the form of a finite sum and do not apply to the *streaming data setting* when each data point is used only once. In this work, we obtain guarantees for both the risk minimization problem and the empirical risk minimization, and our results apply to the streaming data setting. Also, the convergence guarantees provided in Xu et al. (2018) depend on a spectral gap–related parameter that is not provided explicitly, whereas all our results are explicit, and this allows us to have explicit performance comparisons between the upper bounds of SGLD and SGHMC algorithms.

We also note that underdamped Langevin MCMC (also known as Hamiltonian MCMC) and its practical applications are also analyzed further in a number of recent works (see, e.g., Betancourt et al. 2014, 2017; Betancourt 2017; Lee and Vempala 2018; Mangoubi et al. 2018). In particular, Mangoubi et al. (2018) provide a characterization of the conductance of Hamiltonian Monte Carlo (HMC) in continuous time using Liouville's theorem, and invoking Cheeger's inequality, they obtain upper and lower bounds on the spectral gap of HMC in continuous time. Although the formula provided in Mangoubi et al. (2018) for the conductance of HMC is elegant, it is not an explicit formula. In our analysis, our focus is to obtain performance bounds with explicit constants, and therefore, we build on the coupling techniques of Eberle et al. (2019), which leads to explicit constants for the class of problems we consider.

We also note that Mangoubi et al. (2018) consider sampling from the target distribution $\frac{1}{2}\mathcal{N}(-1,\sigma^2) + \frac{1}{2}\mathcal{N}(1,\sigma^2)$ in dimension one and estimate the spectral gap of HMC in the regime as $\sigma \to 0$. This is a mixture of two Gaussians with the same variance σ^2 centered at –1 and 1, respectively, in which they argue that, for this specific example, HMC does not lead to much

improvement over the random walk approach for sampling. In our paper, our results apply to more general targets that are not necessarily a mixture of Gaussians. However, if we consider sampling from the distribution $\frac{1}{2}\mathcal{N}(-a,\sigma^2)+\frac{1}{2}\mathcal{N}(a,\sigma^2)$ as $a\to\infty$ for σ^2 fixed, Proposition 1 is applicable, and it implies that HMC is more efficient than overdamped Langevin dynamics in terms of dependency to a (which measures the distance between the modes) in the sense that the mixing time is $\mathcal{O}(a)$ in HMC, whereas it is $\mathcal{O}(a^2)$ in random walk. This does not contradict the results of Mangoubi et al. (2018) because we consider different scaling regimes: we fix $\sigma>0$ and let $a\to\infty$, whereas Mangoubi et al. (2018) fix $a=1$ and let $\sigma\to 0$.

There are also some connections of our work to existing momentum-based optimization algorithms. More specifically, if the term with $dB(t)$ involving the Brownian noise is removed in the underdamped SDE (5) and (6), this results in a second-order ordinary differential equation (ODE) in $X(t)$. Momentum-based algorithms for strongly convex objectives, such as Polyak's heavy ball method as well as Nesterov's accelerated gradient method, both can be viewed as (alternative) discretizations of this ODE (see, e.g., Polyak 1987, Su et al. 2014, Wilson et al. 2016, Shi et al. 2018). It is known (Su et al. 2014, Wilson et al. 2016, Shi et al. 2018) that Nesterov's accelerated gradient method tracks this second order ODE (also referred to as Nesterov's ODE in the literature), whereas the first order nonaccelerated methods, such as the classical gradient descent, are known to track a first order ODE in $X(t)$ called the *gradient flow* dynamics. Furthermore, existing analysis shows that Nesterov's ODE converges to its equilibrium faster (in time) than the first order gradient flow ODE in terms of upper bounds, and this speed-up is also inherited by the discretized dynamics. Roughly speaking, our results can be interpreted as the analogue of these results in the nonconvex optimization setting in which we deal with SDEs instead of ODEs building on the theory of Markov processes and show that SGHMC tracks the second order (underdamped) Langevin SDE closely and inherits its favorable convergence guarantees (in terms of upper bounds on the expected suboptimality) compared with that of overdamped Langevin SDE.

Acceleration of first order gradient or stochastic gradient methods and their variance-reduced versions for finding a local stationary point (a point with a gradient less than ε in norm) are also studied in the literature (see, e.g., Nesterov 1983, Allen-Zhu and Hazan 2016, Ghadimi and Lan 2016, Carmon et al. 2018, Jofré and Thompson 2019). It is also shown that, under some assumptions, momentum-based accelerated methods can escape saddle points faster (see, e.g., Liu et al. 2018, O'Neill and Wright 2019). In contrast, in this work, our focus is obtaining performance guarantees for convergence to global minimizers instead.

2. Preliminaries and Assumptions

In our analysis, we use the following two-Wasserstein distance: for any two probability measures ν_1,ν_2 on $\mathbb{R}^{2d}$, we define

$$\mathcal{W}_2(\nu_1,\nu_2)=\left(\inf_{Y_1\sim\nu_1,Y_2\sim\nu_2}\mathbb{E}\left[\|Y_1-Y_2\|^2\right]\right)^{1/2},$$

where $\|\cdot\|$ is the usual Euclidean norm, ν_1,ν_2 are two Borel probability measures on $\mathbb{R}^{2d}$ with finite second moments, and the infimum is taken over all random couples (Y_1, Y_2) taking values in $\mathbb{R}^{2d}\times\mathbb{R}^{2d}$ with marginals $Y_1\sim\nu_1, Y_2\sim\nu_2$ (see, e.g., Villani 2008). We let $C^1(\mathbb{R}^d)$ denote the set of continuously differentiable functions on $\mathbb{R}^d$ and $L^2(\pi_{\mathbf{z}})$ denote the space of square-integrable functions on $\mathbb{R}^d$ with respect to the measure $\pi_{\mathbf{z}}$.

We first state the assumptions used in this paper in Assumption 1. Note that we do not assume the component functions $f(x, z)$ to be convex; they can be nonconvex. The first assumption of nonnegativity of f can be assumed without loss of generality by subtracting a constant and shifting the coordinate system as long as f is bounded below. The second assumption of Lipschitz gradients is, in general, unavoidable for discretized Langevin algorithms to be convergent (see, e.g., Mattingly et al. 2002), and the third assumption is known as the *dissipativity condition* (see, e.g., Hale 1988) and is standard in the literature to ensure the convergence of Langevin diffusions to the stationary distribution (see, e.g., Mattingly et al. 2002, Raginsky et al. 2017, Eberle et al. 2019). The fourth assumption is regarding the amount of noise present in the gradient estimates and allows not only constant variance noise, but allows the noise variance to grow with the norm of the iterates (which is the typical situation in mini-batch methods in stochastic gradient methods; see, e.g., Raginsky et al. 2017). Finally, the fifth assumption is a mild assumption saying that the initial distribution μ_0 for the SGHMC dynamics should have a reasonable decay rate of the tails to ensure convergence to the stationary distribution. For instance, if the algorithm is started from any arbitrary point $(x_0,v_0)\in\mathbb{R}^{2d}$, then the Dirac measure $\mu_0(dx,dv)=\delta_{(x_0,v_0)}(dx,dv)$ works. If the initial distribution $\mu_0(dx,dv)$ is supported on a Euclidean ball with radius being some universal constant, it also works. Similar assumptions on the initial distribution μ_0 are also necessary to achieve convergence to a stationary measure in continuous-time underdamped dynamics as well (see, e.g., Hérau and Nier 2004).

Assumption 1. *We impose the following assumptions*:

i. *The function f is continuously differentiable, takes nonnegative real values, and there exist constants $A_0,B\geq 0$ so that*

$$|f(0,z)|\leq A_0,\qquad \|\nabla f(0,z)\|\leq B,$$

for any $z \in \mathcal{Z}$.

ii. *For each* $z \in \mathcal{Z}$, *the function* $f(\cdot, z)$ *is* M*-smooth:*

$$\|\nabla f(w,z) - \nabla f(v,z)\| \le M\|w - v\|.$$

iii. *For each* $z \in \mathcal{Z}$, *the function* $f(\cdot, z)$ *is* (m, b)*-dissipative:*

$$\langle x, \nabla f(x,z)\rangle \ge m\|x\|^2 - b\,.$$

iv. *There exists a constant* $\delta \in [0,1)$ *such that, for every* $\mathbf{z}$,

$$\mathbb{E}[\|g(x,U_{\mathbf{z}}) - \nabla F_{\mathbf{z}}(x)\|^2] \le 2\delta(M^2\|x\|^2 + B^2)\,.$$

v. *The probability law* μ_0 *of the initial state* (X_0, V_0) *satisfies*

$$\int_{\mathbb{R}^{2d}} e^{\alpha \mathcal{V}(x,v)}\mu_0(dx,dv) < \infty\,,$$

where $\mathcal{V}$ *is a Lyapunov function to be used repeatedly for the rest of the paper:*

$$\mathcal{V}(x,v) := \beta F_{\mathbf{z}}(x) + \frac{\beta}{4}\gamma^2(\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda\|x\|^2)\,, \tag{13}$$

and γ *is the friction coefficient as in* (5), λ *is a positive constant less than* $\min(1/4, m/(M + \gamma^2/2))$, *and* $\alpha = \lambda(1 - 2\lambda)/12$.

We note that the Lyapunov function $\mathcal{V}$ is used in Eberle et al. (2019) to study the rate of convergence to equilibrium for underdamped Langevin diffusion, which itself is motivated by, for example, Mattingly et al. (2002). It follows from these assumptions (applying Lemma EC.9) that there exists a constant $A \in (0,\infty)$ so that

$$x \cdot \nabla F_{\mathbf{z}}(x) \ge m\|x\|^2 - b \ge 2\lambda(F_{\mathbf{z}}(x) + \gamma^2\|x\|^2/4) - 2A/\beta\,. \tag{14}$$

This drift condition, which is used later, guarantees the stability and the existence of Lyapunov function $\mathcal{V}$ for the underdamped Langevin diffusion in (5) and (6); see Eberle et al. (2019).

3. Main Results for the SGHMC1 Algorithm

Our first result shows SGHMC1 iterates (X_k, V_k) in (8) and (9) track the underdamped Langevin SDE in the sense that the expectation of the empirical risk $F_{\mathbf{z}}$ with respect to the probability law of (X_k, V_k) conditional on the sample $\mathbf{z}$, denoted by $\mu_{k,\mathbf{z}}$, and the stationary distribution $\pi_{\mathbf{z}}$ of the underdamped SDE is small when k is large enough. The difference in expectations decomposes as a sum of two terms $\mathcal{J}_0(\mathbf{z},\varepsilon)$ and $\mathcal{J}_1(\varepsilon)$ although the former term quantifies the dependency on the initialization and the data set $\mathbf{z}$, whereas the latter term is controlled by the discretization error and the amount of noise in the gradients, which depends on the parameter δ. We also note that the parameter μ_* (see Table 1) in our bounds governs the speed of convergence to the equilibrium of the underdamped Langevin diffusion.

Theorem 1. *Consider the SGHMC1 iterates* (X_k, V_k) *defined by the recursion* (8) *and* (9) *from the initial state* (X_0, V_0), *which has the law* μ_0. *If Assumption* 1 *is satisfied, then for* $\beta, \varepsilon > 0$, *we have*

$$\begin{aligned}\left|\mathbb{E}F_{\mathbf{z}}(X_k) - \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X))\right| &= \left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\mu_{k,\mathbf{z}}(dx,dv)\right.\\ &\quad\left. - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv)\right|\\ &\le \mathcal{J}_0(\mathbf{z},\varepsilon) + \mathcal{J}_1(\varepsilon)\,,\end{aligned}$$

where

$$\mathcal{J}_0(\mathbf{z},\varepsilon) := (M\sigma + B)\cdot C\sqrt{\mathcal{H}_\rho(\mu_0,\pi_{\mathbf{z}})}\cdot\varepsilon, \tag{15}$$

$$\mathcal{J}_1(\varepsilon) := (M\sigma + B)\cdot\left(\left(\frac{C_0}{\mu_*^{3/2}}(\log(1/\varepsilon))^{3/2}\delta^{1/4} + \frac{C_1}{\mu_*^{3/2}}\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))} + \frac{C_2}{\mu_*}\frac{\varepsilon^2}{(\log(1/\varepsilon))^2}\right), \tag{16}$$

with σ *defined by (EC.20) provided that*

$$\eta \le \min\left\{\left(\frac{\varepsilon}{(\log(1/\varepsilon))^{3/2}}\right)^4, 1, \frac{\gamma}{K_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2K_1}, \frac{2}{\gamma\lambda}\right\}, \tag{17}$$

and

$$k\eta = \frac{1}{\mu_*}\log\left(\frac{1}{\varepsilon}\right) \ge e. \tag{18}$$

Here, $\mathcal{H}_\rho$ *is a semimetric for probability distributions defined by (EC.12). All the constants are made explicit and are summarized in Table* 1.

The proof of Theorem 1 is presented in detail in Section EC.1 in the e-companion. In the following sections, we discuss that this theorem combined with some basic properties of the equilibrium distribution $\pi_{\mathbf{z}}$ leads to a number of results that provide performance guarantees for both the empirical and population risk minimization.

3.1. Performance Bound for the Empirical Risk Minimization

In order to obtain guarantees for the empirical risk given in (3), in light of Theorem 1, one has to control the quantity

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x)\,,$$

which is a measure of how much the $x-$ marginal of the equilibrium distribution $\pi_{\mathbf{z}}$ concentrates around a

Table 1. Summary of the Constants and Where They Are Defined in the Text

Constants	Source
$C_x^c = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv)+\frac{(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}$, $C_v^c = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv)+\frac{(d+A)}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}$	(EC.1), (EC.2)
$K_1 = \max\left\{\frac{32M^2\left(\frac{1}{2}+\gamma+\delta\right)}{(1-2\lambda)\beta\gamma^2}, \frac{8\left(\frac{1}{2}M+\frac{1}{4}\gamma^2-\frac{1}{4}\gamma^2\lambda+\gamma\right)}{\beta(1-2\lambda)}\right\}$	(EC.3)
$K_2 = B^2(1+2\gamma+2\delta)$	(EC.4)
$C_x^d = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv)+\frac{4(d+A)}{\lambda}}{\frac{1}{8}(1-2\lambda)\beta\gamma^2}$, $C_v^d = \frac{\int_{\mathbb{R}^{2d}} \mathcal{V}(x,v)\mu_0(dx,dv)+\frac{4(d+A)}{\lambda}}{\frac{\beta}{4}(1-2\lambda)}$	(EC.5), (EC.6)
$\sigma = \max\left\{\sqrt{C_x^c}, \sqrt{C_x^d}\right\} = \sqrt{C_x^d}$	(EC.20)
$C_0 = \hat{\gamma}\cdot\left(\left(M^2C_x^d+B^2\right)\beta/\gamma + \sqrt{(M^2C_x^d+B^2)\beta/\gamma}\right)^{1/2}$	(EC.8)
$C_1 = \hat{\gamma}\cdot\left(\beta M^2(C_2)^2/(2\gamma) + \sqrt{\beta M^2(C_2)^2/(2\gamma)}\right)^{1/2}$	(EC.9)
$C_2 = \left(2\gamma^2C_v^d + (4+2\delta)\left(M^2C_x^d+B^2\right) + 2\gamma\beta^{-1}\right)^{1/2}$	(EC.10)
$\hat{\gamma} = \frac{2\sqrt{2}}{\sqrt{\alpha_0}}\left(\frac{5}{2} + \log\left(\int_{\mathbb{R}^{2d}} e^{\frac{1}{4}\alpha\mathcal{V}(x,v)}\mu_0(dx,dv) + \frac{1}{4}e^{\frac{\alpha(d+A)}{3\lambda}}\alpha\gamma(d+A)\right)\right)^{1/2}$	(EC.11)
$\alpha_0 = \frac{\alpha(1-2\lambda)\beta\gamma^2}{64+32\gamma^2}$, $\alpha = \frac{\lambda(1-2\lambda)}{12}$	(EC.7)
$\mu_* = \frac{\gamma}{768}\min\left\{\lambda M\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}M\gamma^{-2}, \Lambda^{1/2}e^{-\Lambda}\right\}$	(EC.13)
$C = \frac{(1+\gamma)\sqrt{2}e^{1+\frac{\Lambda}{2}}}{\min\{1,\alpha_1\}}\sqrt{\max\{1, 4(1+2\alpha_1+2\alpha_1^2)(d+A)\beta^{-1}\gamma^{-1}\mu_*^{-1}/\min\{1,R_1\}\}}$	(EC.14)
$\Lambda = \frac{12}{5}(1+2\alpha_1+2\alpha_1^2)(d+A)M\gamma^{-2}\lambda^{-1}(1-2\lambda)^{-1}$, $\alpha_1 = (1+\Lambda^{-1})M\gamma^{-2}$	(EC.15)
$\varepsilon_1 = 4\gamma^{-1}\mu_*/(d+A)$	(EC.16)
$R_1 = 4\cdot(6/5)^{1/2}(1+2\alpha_1+2\alpha_1^2)^{1/2}(d+A)^{1/2}\beta^{-1/2}\gamma^{-1}(\lambda-2\lambda^2)^{-1/2}$	(EC.17)
$\overline{\mathcal{H}}_\rho(\mu_0) = R_1 + R_1\varepsilon_1\max\left\{M+\frac{1}{2}\beta\gamma^2, \frac{3}{4}\beta\right\}\|(x,v)\|^2_{L^2(\mu_0)} + R_1\varepsilon_1\left(M+\frac{1}{2}\beta\gamma^2\right)\frac{b+d/\beta}{m} + R_1\varepsilon_1\frac{3}{4}d + 2R_1\varepsilon_1\left(\beta A_0 + \frac{\beta B^2}{2M}\right)$	(EC.18)

global minimizer of the empirical risk. As β goes to infinity, it can be verified that this quantity goes to zero. For finite β, Raginsky et al. (2017) (see proposition 11) derives an explicit bound of the form

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \le \mathcal{J}_2 := \frac{d}{2\beta}\log\left(\frac{eM}{m}\left(\frac{b\beta}{d}+1\right)\right), \tag{19}$$

(which is also provided in the e-companion for the sake of completeness, see Lemma EC.12). This combined with Theorem 1 immediately leads to the following performance bound for the empirical risk minimization. The proof is omitted.

Corollary 1 (Empirical Risk Minimization with SGHMC1). *Under the setting of Theorem 1, the empirical risk minimization problem admits the performance bounds*

$$\mathbb{E}F_{\mathbf{z}}(X_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \le \mathcal{J}_0(\varepsilon,\mathbf{z}) + \mathcal{J}_1(\varepsilon) + \mathcal{J}_2\,,$$

provided that Conditions (17) and (18) hold in which the terms $\mathcal{J}_0(\mathbf{z},\varepsilon)$, $\mathcal{J}_1(\varepsilon)$ and $\mathcal{J}_2$ are defined by (15), (16), and (19), respectively.

3.2. Performance Bound for the Population Risk Minimization

By exploiting the fact that the $x-$ marginal of the invariant distribution for the underdamped dynamics is the same as it is in the overdamped case, it can be shown that the generalization error $F(X_k) - F_{\mathbf{Z}}(X_k)$ is no worse than that of the available bounds for SGLD given in Raginsky et al. (2017), and therefore, we have the following corollary. A more detailed proof is given in Section EC.1.

Corollary 2 (Population Risk Minimization with SGHMC1). *Under the setting of Theorem 1, the expected population risk of X_k (the iterates in (9)) is bounded by*

$$\mathbb{E}F(X_k) - F^* \le \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon) + \mathcal{J}_2 + \mathcal{J}_3(n)\,,$$

with

$$\overline{\mathcal{J}}_0(\varepsilon) := (M\sigma+B)\cdot C\cdot\sqrt{\overline{\mathcal{H}}_\rho(\mu_0)}\cdot\varepsilon, \tag{20}$$

$$\mathcal{J}_3(n) := \frac{4\beta c_{LS}}{n}\left(\frac{M^2}{m}(b+d/\beta)+B^2\right), \tag{21}$$

where σ is defined by (EC.20); $\overline{\mathcal{H}}_\rho(\mu_0)$ is defined by (EC.18); $\mathcal{J}_1(\varepsilon)$ and $\mathcal{J}_2$ are defined by (16) and (19), respectively; and c_{LS} is a constant satisfying

$$c_{LS} \le \frac{2m^2+8M^2}{m^2M\beta} + \frac{1}{\lambda_*}\left(\frac{6M(d+\beta)}{m}+2\right),$$

and λ_* is the uniform spectral gap for overdamped Langevin dynamics:[3]

$$\lambda_* := \inf_{\mathbf{z}\in\mathcal{Z}^n} \inf\left\{\frac{\beta^{-1}\int_{\mathbb{R}^d}\|\nabla g\|^2 d\pi_{\mathbf{z}}}{\int_{\mathbb{R}^d} g^2 d\pi_{\mathbf{z}}} : g\in C^1(\mathbb{R}^d)\cap L^2(\pi_{\mathbf{z}}),\ g\neq 0, \int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0\right\}. \tag{22}$$

3.3. Generalization Error of SGHMC1 in the One-Pass Regime

Because the predictor X_k is random, it is natural to consider the expected generalization error $\mathbb{E}F(X_k) - \mathbb{E}F_{\mathbf{Z}}(X_k)$ (see, e.g., Hardt et al. 2016) that admits the decomposition

$$\begin{aligned}\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k) &= (\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)) \\ &\quad + (\mathbb{E}F_{\mathbf{Z}}(X^\pi) - \mathbb{E}F(X^\pi)) \\ &\quad + (\mathbb{E}F(X^\pi) - \mathbb{E}F(X_k)),\end{aligned} \tag{23}$$

where X^π is the Gibbs output; that is, its distribution conditional on $\mathbf{Z}=\mathbf{z}$ is given by $\pi_{\mathbf{z}}$. If every sample is used once, that is, if only one pass is made over the data set, then the second term in (23) disappears. As a consequence, the generalization error is controlled by the bound

$$\begin{aligned}|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k)| &\le |\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)| \\ &\quad + |\mathbb{E}F(X^\pi) - \mathbb{E}F(X_k)|.\end{aligned} \tag{24}$$

The following result provides a bound on this quantity. The proof is similar to the proof of Theorem 1 and its corollaries and, hence, omitted.

Theorem 2 (Generalization Error of SGHMC1). *Under the setting of Theorem 1, we have*

$$|\mathbb{E}F(X_k) - \mathbb{E}F(X^\pi)| \le \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon),$$

$$|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)| \le \overline{\mathcal{J}}_0(\varepsilon) + \mathcal{J}_1(\varepsilon),$$

provided that (17) and (18) hold when X^π is the output of the underdamped Langevin dynamics, that is, its distribution conditional on $\mathbf{Z}=\mathbf{z}$ is given by $\pi_{\mathbf{z}}$ and $\overline{\mathcal{J}}_0(\varepsilon)$ is defined by (20). Then, it follows from (24) that, if each data point is used once, the expected generalization error satisfies

$$|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k)| \le 2\overline{\mathcal{J}}_0(\varepsilon) + 2\mathcal{J}_1(\varepsilon).$$

4. Main Results for the SGHMC2 Algorithm

Recall the SGHMC2 algorithm $(\hat{X}_k, \hat{V}_k)$ defined in (10) and (11), and denote the probability law of $(\hat{X}_k, \hat{V}_k)$ conditional on the sample $\mathbf{z}$ by $\hat{\mu}_{k,\mathbf{z}}(dx,dv)$. Similar to our analysis for SGHMC1, we can derive similar performance guarantees for SGHMC2 in terms of empirical risk, population risk, and the generalization error. The main difference is that the term $\mathcal{J}_1(\varepsilon)$ is controlled by the accuracy of the discretization and has to be replaced by another term $\hat{\mathcal{J}}_1(\varepsilon)$ as the SGHMC2 algorithm is based on an alternative discretization. In particular, the performance bounds we get for SGHMC2 are tighter than SGHMC1 as is elaborated further in Section 5.

Theorem 3. *Consider the SGHMC2 iterates $(\hat{X}_k, \hat{V}_k)$ defined by the recursion (10) and (11) from the initial state (X_0, V_0) which has the law μ_0. If Assumption 1 is satisfied, then for $\beta, \varepsilon > 0$, we have*

$$\begin{aligned}\left|\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X))\right| &= \left|\int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\hat{\mu}_{k,\mathbf{z}}(dx,dv) \right. \\ &\quad \left. - \int_{\mathbb{R}^d\times\mathbb{R}^d} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv)\right| \\ &\le \mathcal{J}_0(\mathbf{z},\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon),\end{aligned}$$

where $\mathcal{J}_0(\mathbf{z},\varepsilon)$ is defined in (15) and

$$\hat{\mathcal{J}}_1(\varepsilon) := (M\sigma + B)\cdot\left(\frac{C_0}{\sqrt{\mu_*}}\sqrt{\log(1/\varepsilon)}\delta^{1/4} + \frac{\hat{C}_1}{\sqrt{\mu_*}}\varepsilon\right) \sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}, \tag{25}$$

with σ defined by (EC.20) provided that

$$\eta \le \min\left\{\left(\frac{\varepsilon}{\sqrt{\log(1/\varepsilon)}}\right)^2, 1, \frac{\gamma}{\hat{K}_2}(d/\beta + A/\beta), \frac{\gamma\lambda}{2\hat{K}_1}, \frac{2}{\gamma\lambda}\right\}, \tag{26}$$

and

$$k\eta = \frac{1}{\mu_*}\log\left(\frac{1}{\varepsilon}\right) \ge e. \tag{27}$$

Here, $\mathcal{H}_\rho$ is a semimetric for probability distributions defined by (EC.12). All the constants are made explicit and are summarized in Tables 1 and 2.

The proof of Theorem 3 is given in Section EC.2. Relying on Theorem 3, one can readily derive the following result on the performance bound for the empirical risk minimization with the SGHMC2 algorithm. The proof follows a similar argument as discussed in Section 3.1 and is omitted.

Table 2. Summary of the Constants and Where They Are Defined in the Text

Constants	Source
$\hat{K}_1 = K_1 + Q_1\frac{4}{1-2\lambda} + Q_2\frac{8}{(1-2\lambda)\gamma^2}$	(EC.22)
$\hat{K}_2 = K_2 + Q_3$	(EC.23)
$Q_1 = \frac{1}{2}c_0((5M+4-2\gamma+(c_0+\gamma^2)) + (1+\gamma)\left(\frac{5}{2}+c_0(1+\gamma)\right) + 2\gamma^2\lambda)$	(EC.24)
$Q_2 = \frac{1}{2}c_0[((1+\gamma)\left(c_0(1+\gamma)+\frac{5}{2}\right) + c_0 + 2 + \lambda\gamma^2 + 2(Mc_0+M+1))(2(1+\delta)M^2) + \left(2M^2+\gamma^2\lambda+\frac{3}{2}\gamma^2(1+\gamma)\right)$	(EC.25)
$Q_3 = c_0((1+\gamma)\left(c_0(1+\gamma)+\frac{5}{2}\right) + c_0 + 2 + \lambda\gamma^2 + 2(Mc_0+M+1))(1+\delta)B^2 + c_0B^2 + \frac{1}{2}\gamma^3\beta^{-1}c_{22} + \gamma^2\beta^{-1}c_{12} + M\gamma\beta^{-1}c_{22}$	(EC.26)
$c_0 = 1+\gamma^2, \quad c_{12} = \frac{d}{2}, \quad c_{22} = \frac{d}{3}$	(EC.27)
$\hat{C}_1 = \hat{\gamma}\cdot\left(\frac{3\beta M^2}{2\gamma}\left(C_v^d + (2(1+\delta)M^2C_x^d + 2(1+\delta)B^2) + \frac{2d\gamma\beta^{-1}}{3}\right) + \sqrt{\frac{3\beta M^2}{2\gamma}\left(C_v^d + (2(1+\delta)M^2C_x^d + 2(1+\delta)B^2) + \frac{2d\gamma\beta^{-1}}{3}\right)}\right)^{1/2}$	(EC.29)

Corollary 3 (Empirical Risk Minimization with SGHMC2). *Under the setting of Theorem 3, the empirical risk minimization problem admits the performance bounds*

$$\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \le \mathcal{J}_0(\mathbf{z},\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon) + \mathcal{J}_2,$$

provided that Conditions (26) *and* (27) *hold in which the terms* $\mathcal{J}_0(\mathbf{z},\varepsilon)$, $\hat{\mathcal{J}}_1(\varepsilon)$ *and* $\mathcal{J}_2$ *are defined by* (15), (25), *and* (19), *respectively.*

Next, we present the performance bound for the population risk minimization with the SGHMC2 algorithm. Similarly to in Section 3.2, to control the population risk during SGHMC2 iterations, one needs to control the difference between the finite sample size Problem (2) and the original Problem (1) in addition to the empirical risk. This leads to the following result. The details of the proof are given in Section EC.2.

Corollary 4 (Population Risk Minimization with SGHMC2). *Under the setting of Theorem 3, the expected population risk of* $\hat{X}_k$ *(the iterates in* (11)*) is bounded by*

$$\mathbb{E}F(\hat{X}_k) - F^* \le \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon) + \mathcal{J}_2 + \mathcal{J}_3(n),$$

where $\overline{\mathcal{J}}_0(\varepsilon)$, $\hat{\mathcal{J}}_1(\varepsilon)$, $\mathcal{J}_2$, $\mathcal{J}_3(n)$ *are defined in* (20), (25), (19), *and* (21).

Finally, we present a result on the generalization error of the SGHMC2 algorithm in the one-pass regime. The proof follows from Theorem 3 and the discussion for the SGHMC1 algorithm in Section 3.3 and, hence, is omitted.

Theorem 4 (Generalization Error of SGHMC2). *Under the setting of Theorem 3, we have*

$$\left|\mathbb{E}F(\hat{X}_k) - \mathbb{E}F(X^\pi)\right| \le \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon),$$
$$\left|\mathbb{E}F_{\mathbf{Z}}(\hat{X}_k) - \mathbb{E}F_{\mathbf{Z}}(X^\pi)\right| \le \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon),$$

provided that (26) *and* (27) *hold, where* X^π *is the output of the underdamped Langevin dynamics; that is, its distribution conditional on* $\mathbf{Z}=\mathbf{z}$ *is given by* $\pi_{\mathbf{z}}$ *and* $\overline{\mathcal{J}}_0(\varepsilon)$ *is defined by* (20). *Then, it follows from* (24) *that, if each data point is used once, the expected generalization error satisfies*

$$|\mathbb{E}F_{\mathbf{Z}}(\hat{X}_k) - \mathbb{E}F(\hat{X}_k)| \le 2\overline{\mathcal{J}}_0(\varepsilon) + 2\hat{\mathcal{J}}_1(\varepsilon).$$

5. Performance Comparison with Respect to the SGLD Algorithm

In this section, we compare our performance bounds for SGHMC1 and SGHMC2 to SGLD. We use the notations $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$ to give explicit dependence on the parameters d,β,μ_*, but it hides factors that depend (at worst polynomially) on other parameters m,M,B,λ,γ,b and A_0. Without loss of generality, we assume here the initial distribution $\mu_0(dx,dv)$ is supported on a Euclidean ball with radius being some universal constant for the simplicity of performance comparison.

5.1. Generalization Error in the One-Pass Setting

A consequence of Theorem 2 is that the generalization error of the SGHMC1 iterates $|\mathbb{E}F_{\mathbf{Z}}(X_k) - \mathbb{E}F(X_k)|$ in the one-pass setting satisfy

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon + \frac{(d+\beta)^{3/2}}{\beta(\mu_*)^{3/2}}\left((\log(1/\varepsilon))^{3/2}\delta^{1/4}+\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))} + \frac{d+\beta}{\beta}\frac{\varepsilon^2}{\mu_*(\log(1/\varepsilon))^2}\right), \tag{28}$$

for $k = K_{SGHMC1} := \tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^4}\log^7(1/\varepsilon)\right)$ iterations, and similarly, Theorem 4 implies the generalization error of the SGHMC2 iterates $|\mathbb{E}F_{\mathbf{Z}}(\hat{X}_k) - \mathbb{E}F(\hat{X}_k)|$ in the one-pass setting satisfy

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon + \frac{(d+\beta)^{3/2}}{\beta\sqrt{\mu_*}}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}\right), \tag{29}$$

for $k = K_{SGHMC2} := \tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^2}\log^2(1/\varepsilon)\right)$ iterations (see the discussion in Section EC.7 for details). On the other

hand, the results of theorem 1 in Raginsky et al. (2017) imply that the generalization error for the SGLD algorithm after K_{SGLD} iterations in the one-pass setting scales as

$$\tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*}\left(\delta^{1/4}\log(1/\varepsilon)+\varepsilon\right)\right) \quad \text{for}$$
$$K_{SGLD}=\tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_*\varepsilon^4}\log^5(1/\varepsilon)\right). \tag{30}$$

The constants λ_* (see (22)) and μ_* (see Table 1) are exponentially small in both β and d in the worst case, but under some extra assumptions, the dependency on d can be polynomial (see, e.g., Cheng et al. 2018b) although the exponential dependence on β is unavoidable in the presence of multiple minima in general (see Bovier et al. 2005). One can readily see that K_{SGHMC2} has better dependency on ϵ than K_{SGHMC1} and infer from (28) and (29) that the performance of SGHMC2 is better than SGHMC1. Hence, in the rest of the section, we only focus on the comparison between SGHMC2 and SGLD.

We see that the generalization error for SGHMC2 (29) is bounded by

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\sqrt{\log\log(1/\varepsilon)}\right), \tag{31}$$

as μ_* is small, and if we ignore the $\sqrt{\log\log(1/\varepsilon)}$ factor,[4] then we get

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\right) \quad \text{for}$$
$$K_{SGHMC2}=\tilde{\Omega}\left(\frac{1}{\mu_*\varepsilon^2}\log^2(1/\varepsilon)\right) \tag{32}$$

iterations of the SGHMC2 algorithm, whereas the corresponding bound for SGLD from Raginsky et al. (2017), theorem 1, is

$$\tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*}\left(\log(1/\varepsilon)\delta^{1/4}+\varepsilon\right)\right) \quad \text{for}$$
$$K_{SGLD}=\tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_*\varepsilon^4}\log^5(1/\varepsilon)\right) \tag{33}$$

iterations of the SGLD algorithm. Note that K_{SGHMC2} and K_{SGLD} do not have the same dependency on ε up to log factors (the former scales with ε as $\log^2(1/\varepsilon)\varepsilon^{-2}$ and the latter $\log^5(1/\varepsilon)\varepsilon^{-4}$), and this improvement on ε dependency is due to better diffusion approximation of SGHMC2 (see Lemma EC.6) compared with SGLD and the exponential integrability estimate we have in Lemma EC.2, which improves the estimate in Raginsky et al. (2017), and using the same argument, one can improve the $\log^5(1/\varepsilon)/\varepsilon^4$ term in (33) to $\log^3(1/\varepsilon)/\varepsilon^4$.

To make the comparison with SGLD simpler, we notice that in both Expressions (32) and (33), we see a term scaling with $\delta^{1/4}$ because of the gradient noise level δ (δ is fixed in the one-pass setting), and we fix the error in (32) and (33) without the δ term to be the same order and then compare the number of iterations K_{SGHMC2} and K_{SGLD}. More precisely, given $\hat{\varepsilon}>0$, and we choose $\varepsilon>0$ such that $\frac{(d+\beta)^{3/2}}{\beta\mu_*}\varepsilon=\hat{\varepsilon}$ in (32) so that the generalization error for SGHMC2 is

$$\tilde{\mathcal{O}}\left(\hat{\varepsilon}+\frac{(d+\beta)^{3/2}}{\beta\mu_*}\sqrt{\log\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*\hat{\varepsilon}}\right)}\delta^{1/4}\right) \quad \text{for}$$
$$K_{SGHMC2}=\tilde{\Omega}\left(\frac{(d+\beta)^3}{\beta^2\mu_*^3\hat{\varepsilon}^2}\log^2\left(\frac{(d+\beta)^{3/2}}{\beta\mu_*\hat{\varepsilon}}\right)\right). \tag{34}$$

Similarly, the generalization error for SGLD is

$$\tilde{\mathcal{O}}\left(\hat{\varepsilon}+\frac{\beta(\beta+d)^2}{\lambda_*}\log\left(\frac{\beta(\beta+d)^2}{\lambda_*\hat{\varepsilon}}\right)\delta^{1/4}\right) \quad \text{for}$$
$$K_{SGLD}=\tilde{\Omega}\left(\frac{\beta^5(d+\beta)^9}{\lambda_*^5\hat{\varepsilon}^4}\log^5\left(\frac{\beta(\beta+d)^2}{\lambda_*\hat{\varepsilon}}\right)\right). \tag{35}$$

When λ_* and μ_* are on the same order or μ_* is larger, because typically $\beta\geq 1$, the term involving δ in the generalization error for SGHMC2 is (smaller) better than the counterpart for SGLD, and this is guaranteed to be achieved in a fewer number of iterations ignoring the log factors and universal constants for K_{SGHMC2} in (34) and K_{SGLD} in (35).

Comparing λ_* and μ_* on arbitrary nonconvex functions seems nontrivial; however, we give a class of nonconvex functions (see Proposition 1 and Example 1), where $\frac{1}{\mu_*}=\tilde{\mathcal{O}}\left(\sqrt{\frac{1}{\lambda_*}}\right)$. For this class, we can infer from (34) that K_{SGHMC2} has a dependency $1/\mu_*^3=\tilde{\mathcal{O}}(1/\lambda_*^{3/2})$, which is much smaller in contrast to $1/\lambda_*^5$ for K_{SGLD} in (35).

5.2. Empirical Risk Minimization

The empirical risk minimization bound given in Corollary 3 has an additional term $\mathcal{J}_2$ compared with the $\overline{\mathcal{J}}_0(\varepsilon)$ and $\hat{\mathcal{J}}_1(\varepsilon)$ terms appearing in the one-pass generalization bounds. Note also that $\mathcal{J}_0(\mathbf{z},\varepsilon)\leq\overline{\mathcal{J}}_0(\varepsilon)$. As a consequence, the SGHMC2 algorithm has expected empirical risk

$$\tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\mu_*\beta^{5/4}}\varepsilon+\frac{(d+\beta)^{3/2}}{\beta\sqrt{\mu_*}}\left(\sqrt{\log(1/\varepsilon)}\delta^{1/4}+\varepsilon\right)\right.$$
$$\left.\sqrt{\log(\mu_*^{-1}\log(\varepsilon^{-1}))}+d\cdot\frac{\log(1+\beta)}{\beta}\right), \tag{36}$$

after $K_{SGHMC2} = \tilde{\Omega}\left(\frac{1}{\mu_* \varepsilon^2}\log^2(1/\varepsilon)\right)$ iterations as opposed to

$$\tilde{\mathcal{O}}\left(\frac{\beta(\beta+d)^2}{\lambda_*}\left(\delta^{1/4}\log(1/\varepsilon)+\varepsilon\right)+d\cdot\frac{\log(1+\beta)}{\beta}\right), \quad (37)$$

after $K_{SGLD} = \tilde{\Omega}\left(\frac{\beta(d+\beta)}{\lambda_* \varepsilon^4}\log^5(1/\varepsilon)\right)$ iterations required in Raginsky et al. (2017). Comparing (36) and (37), we see that the last terms are the same. If this term is the dominant term, then the empirical risk upper bounds for SGLD and SGHMC2 are similar except that K_{SGHMC2} can be smaller than K_{SGLD}, for instance, when $\frac{1}{\mu_*} = \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{\lambda_*}}\right)$. Otherwise, if the last term is not the dominant one and can be ignored with respect to other terms, then the performance comparison is similar to the discussion about the generalization bounds (31) and (33).

We next briefly discuss the comparisons of SGHMC2 and SGLD based on the total number of stochastic gradient evaluations (gradient complexity), and we compare with a recent work (Xu et al. 2018), which establishes a faster convergence result and improves the gradient complexity for SGLD in the mini-batch setting compared with Raginsky et al. (2017). Here, the total number of stochastic gradient evaluations of an algorithm is defined as the number of stochastic gradients calculated per iteration (which is equal to the batch size in the mini-batch setting) times the total number of iterations. Xu et al. (2018) show that it suffices to take

$$\hat{K}_{SGLD} = \tilde{\Omega}\left(\frac{d^7}{\hat{\lambda}^5 \hat{\varepsilon}^5}\right) \quad (38)$$

stochastic gradient evaluations to converge to an $\hat{\varepsilon}$ neighborhood of an almost ERM, where $\tilde{\Omega}(\cdot)$ hides some factors in β and $\hat{\lambda}$ is the spectral gap of the discrete overdamped Langevin dynamics, that is, SGLD with zero gradient noise. This improves upon the result in Raginsky et al. (2017), which shows that the same task requires $\tilde{\Omega}\left(\frac{d^{17}}{\lambda_*^9 \hat{\varepsilon}^8}\right)$ stochastic gradient evaluations. Our results show that (see, e.g., (36)), for SGHMC2, it suffices to have

$$\hat{K}_{SGHMC2} = \tilde{\Omega}\left(\frac{d^9}{\mu_*^4 \hat{\varepsilon}^6}\right) \quad (39)$$

stochastic gradient evaluations, ignoring the log factors in the parameters $\hat{\varepsilon}, \mu_*, d$ and hiding factors in β that can be made explicit. To see (39), we infer from (36) that, for fixed precision $\hat{\varepsilon} > 0$ and dimension d, by ignoring the log factors and β, we can choose ε so that $d^{3/2}\varepsilon/\mu_* = \hat{\varepsilon}$ and choose the gradient noise level δ so that $d^{3/2}\delta^{1/4}/\sqrt{\mu_*} = \hat{\varepsilon}$. So the number of SGHMC2 iterations is

$$K_{SGHMC2} = \tilde{\Omega}\left(\frac{1}{\mu_* \varepsilon^2}\right) = \tilde{\Omega}\left(\frac{d^3}{\mu_*^2 \hat{\varepsilon}^2}\right).$$

On the other hand, the mini-batch size to achieve gradient noise level δ is given by $1/\delta$ (see Raginsky et al. 2017), which is equal to $d^6/(\mu_*^2 \hat{\varepsilon}^4)$. Hence, we obtain (39), which is the product of the mini-batch size and number of iterations.

It is hard to compare $\hat{\lambda}$ in (38) and μ_* in (39) in general because $\hat{\lambda}$ is the spectral gap of the discrete overdamped Langevin dynamics (i.e., SGLD with zero gradient noise) without a simple closed-form formula. However, when the step size is small enough, we expect $\hat{\lambda}$ is similar to λ_*, which is the spectral gap of the continuous-time overdamped Langevin diffusion. As a consequence, when the step size η is small enough (which is the case, for instance, when target accuracy $\hat{\varepsilon}$ is small enough), we have $\hat{\lambda} \approx \lambda_*$ and $\frac{1}{\mu_*} = \mathcal{O}\left(\sqrt{\frac{1}{\lambda_*}}\right) = \mathcal{O}\left(\sqrt{\frac{1}{\hat{\lambda}}}\right)$ for the class of nonconvex functions we discuss in Proposition 1 and Example 1. For this class of problems, comparing (38) and (39), we see that we obtain an improvement in the spectral gap parameter (μ_*^4 versus $\hat{\lambda}^5$); however, $\hat{\varepsilon}$ and d dependency of the bound (38) is better than (39).

5.3. Population Risk Minimization

If samples are recycled and multiple passes over the data set are made, then one can see from Corollary 2 that there is an extra term $\mathcal{J}_3$ that needs to be added to the bounds given in (36) and (37). This term satisfies

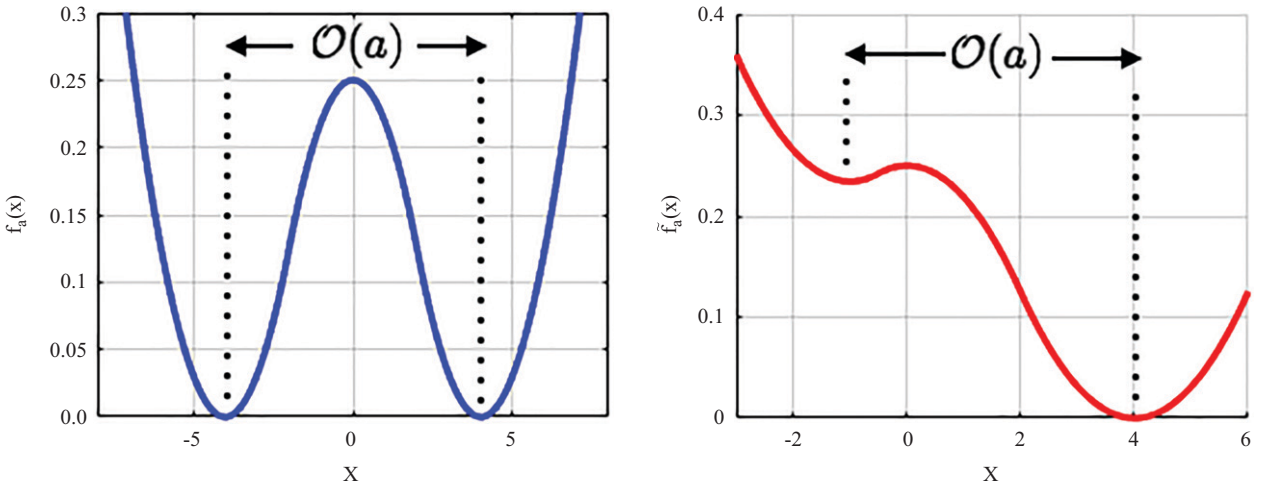
$$\mathcal{J}_3 = \tilde{\mathcal{O}}\left(\frac{(\beta+d)^2}{\lambda_* n}\right).$$

If this term is dominant compared with other terms $\overline{\mathcal{J}_0}, \mathcal{J}_1$ and $\mathcal{J}_2$, for instance, this may happen if the number of samples n is not large enough, and then, the performance guarantees for population risk minimization via SGLD and SGHMC2 are similar. Otherwise, if n is large and β is chosen in a way to keep the $\mathcal{J}_2$ term on the order $\overline{\mathcal{J}_0}$, then similar improvement can be achieved.

5.4. Comparison of λ_* and μ_*

The parameters λ_* (see (22)) and μ_* (see Table 1) govern the convergence rate to the equilibrium of the overdamped and underdamped Langevin SDE, and they can be both exponentially small in dimensions d and in β. They appear naturally in the complexity estimates of the SGHMC2 and SGLD methods as these algorithms can be viewed as discretizations of Langevin SDEs (when the discretization step is small and the

Figure 1. (Color online) The Illustration of the Functions $f_a(x)$ (Left) and $\tilde{f}_a(x)$ (Right) for $a = 4$



gradient noise $\delta = 0$, the discrete dynamics behave similarly to the continuous dynamics). Next, to get further intuition, first, we discuss some toy examples of nonconvex functions in which $\frac{1}{\mu_*} = \mathcal{O}\left(\sqrt{\frac{1}{\lambda_*}}\right)$. For these examples, if the other parameters (β, d, δ) are fixed, then SGHMC2 can lead to an improvement upon the SGLD performance. We then show in Proposition 1 that these examples generalize to a more general class of nonconvex functions.

Example 1. Consider the following symmetric double-well potential in $\mathbb{R}^d$ studied previously in the context of Langevin diffusions (Eberle et al. 2019):

$$f_a(x) = U(x/a) \quad \text{with} \quad U(x) := \begin{cases} \frac{1}{2}(\|x\| - 1)^2 & \text{for} \quad \|x\| \geq \frac{1}{2}, \\ \frac{1}{4} - \frac{\|x\|^2}{2} & \text{for} \quad \|x\| \leq \frac{1}{2}, \end{cases}$$

where $a > 0$ is a scaling parameter that is illustrated in the left panel of Figure 1. For this example, there are two minima that are apart at a distance $\mathcal{R} = \mathcal{O}(a)$. For simplicity, we assume there is only one sample, that is, $\mathbf{z} = (z_1)$ and $F_{\mathbf{z}}(x) = f(x, z_1) = f_a(x)$. We consider the nonconvex optimization Problem (2) with both the SGHMC2 and SGLD algorithms. Eberle et al. (2019) show that $\mu_* \geq \Theta(\frac{1}{a})$ for this example, whereas $\lambda_* \leq \Theta(\frac{1}{a^2})$, making the constants hidden by the Θ explicit. This shows that the contraction rate of the underdamped diffusion μ_* is (faster) larger than that of the overdamped diffusion λ_* by a square root factor when a is large and all the constants can be made explicit. Such results extend to a more general class of nonconvex functions with multiple wells and higher dimensions as long as the gradient of the objective satisfies a growth condition (see examples 1.1 and 1.13 in Eberle et al. (2019) for a further discussion).

For computing an ε-approximate global minimizer of $f_a = f(x, z_1)$ (or, more generally, for a nonconvex problem satisfying Assumption 1), β is chosen large enough so that the stationary measure concentrates around the global minimizer. Using the tight characterization of λ_* from theorem 1.2 in Bovier et al. (2005) for β large, further comparisons with similar conclusions between the rate of convergence to the equilibrium distribution between the underdamped and overdamped dynamics can also be made. For example, consider the nonconvex objective $F_{\mathbf{z}}(x) = \tilde{f}_a(x) = \tilde{U}(x/a)$ instead, illustrated in the right panel of Figure 1 for $a = 4$, where

$$\tilde{U}(x) = \begin{cases} \frac{1}{2}(x-1)^2 & \text{for} \quad x \geq \frac{1}{2}, \\ \frac{1}{4} - \frac{x^2}{2} & \text{for} \quad -\frac{1}{8} \leq x \leq \frac{1}{2}, \\ \frac{1}{2}\left(x + \frac{1}{4}\right)^2 + \frac{15}{64} & \text{for} \quad x \leq -\frac{1}{8}, \end{cases}$$

is the asymmetric double-well potential in dimension one. It follows from Theorem EC.1 (see also Eberle et al. 2019) that the contraction rate satisfies $\mu_* = \Theta(a^{-1})$, whereas it follows from theorem 1.2 in Bovier et al. (2005) that $\lambda_* = \Theta(1/a^2)$. This shows that, when the separation between minima or, alternatively, the scaling factor a is large enough, μ_* is larger than λ_* by a square root factor up to constants.

The behavior in these toy examples can be generalized to more general nonconvex objectives with a finite-sum structure satisfying Assumption 1. Proposition 1 gives a class of functions in which μ_* is on the order of the square root of λ_*. The proof is presented in detail in Section EC.6.

Proposition 1. *Suppose that the functions $f_a(x, z)$ indexed by a satisfies Assumption 1(i)–(iii) with $m = m_1 a^{-2}$, $M =$*

$M_1 a^{-2}$ *and* $B = B_1 a^{-1}$ *for some fixed constants* m_1, M_1, *and* B_1. *Then, we have, as* $a \to \infty$,

$$\lambda_* = \mathcal{O}(a^{-2}), \qquad \mu_* = \Theta(a^{-1}). \tag{40}$$

This result is more general than the previous example. In particular, if $f(x, z)$ satisfies Assumption 1(i)–(iii) with m, M, and B replaced by m_1, M_1, B_1, then $f_a(x,z) := f(x/a, z)$ satisfies Assumption 1(i)–(iii) with $m = m_1 a^{-2}$, $M = M_1 a^{-2}$ and $B = B_1 a^{-1}$. Proposition 1 essentially says that, if we consider the normalized empirical risk objective $F_{\mathbf{z}}(x/a) = \frac{1}{n}\sum_{i=1}^n f(x/a, z_i)$, where a is a (normalization) scaling parameter and $f(x, z)$ satisfies Assumption 1, then for large enough values of a, the empirical risk surface is relatively flat, and the convergence rate of momentum variant SGHMC2 to an ε-neighborhood of the global minimum is governed by the parameter μ_*, which is larger than that of the parameter λ_* of SGLD when a is sufficiently large. This leads to improved performance bounds for SGHMC2 compared with known performance bounds for SGLD.

6. Applications

We note that several nonconvex stochastic optimization problems of interest can satisfy Assumption 1 under appropriate noise assumptions for the underlying data set. For example, LASSO problems with nonconvex regularizers (see, e.g., Hu et al. 2017); nonconvex formulations of the phase retrieval problem around global minimum (see, e.g., Zhang et al. 2017b); or nonconvex stochastic optimization problems defined on a compact set, including but not limited to dictionary learning over the sphere (see, e.g., Sun et al. 2016) and training deep learning models subject to norm constraints in the model parameters (see, e.g., Anil et al. 2019). In this section, we discuss some applications of our results, and we provide two specific examples.

6.1. Binary Linear Classification

In binary linear classification, the aim is to learn a predictive model of the form $\mathbb{P}(Y = 1 | A_{in} = a) = \sigma_c(\langle \tilde{x}, a \rangle)$, where $\tilde{x} \in \mathbb{R}^d$ is a parameter vector to be learned, A_{in} is the input variable (feature vector), Y is the binary output, and $\sigma_c : \mathbb{R} \to [0,1]$ is a threshold function. Binary classification arises in many data-driven applications in operations research from diagnosing patients in healthcare (Wu and Liu 2007) to predicting directions in the stock market (James et al. 2013).

A number of empirical studies demonstrate that nonconvex choices of the σ_c function can often lead to superior classification accuracy and robustness properties compared with convex choices of σ_c, such as the hinge loss (Collobert et al. 2006, Wu and Liu 2007, Chapelle et al. 2009, Nguyen and Sanner 2013). Given access to a data set of input–output pairs $z_i = (a_i, y_i)$, a standard way of estimating $\tilde{x}$ is based on minimizing the *regularized squared loss* over the data set, that is,

$$\min_{x \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (y_i - \sigma_c(\langle x, a_i \rangle))^2 + \frac{\lambda_r}{2}\|x\|^2, \tag{41}$$

where $\lambda_r > 0$ is a regularization parameter that may depend on the number of samples n. By Lagrangian duality, this problem is equivalent to the constrained optimization problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (y_i - \sigma_c(\langle x, a_i \rangle))^2 \quad \text{subject to} \quad \|x\| \le R,$$

for some R, which has also been considered in the literature (see, e.g., Foster et al. 2018, Mei et al. 2018, Wang et al. 2019). For nonconvex $\sigma_c(\cdot)$, this problem is also nonconvex in general. We consider minimizing the objective (41) in the mini-batch setting in which the gradients in SGHMC iterations are estimated from n_b data points sampled with replacement,; that is, the gradient is estimated as

$$g(x, U_{\mathbf{z}}) = \frac{1}{n_b}\sum_{j=1}^{n_b} \nabla f(x, z_j), \tag{42}$$

where z_j are i.i.d. with a uniform distribution over the set of indices $\{1, 2, \ldots, n\}$. We also consider the following assumption for the threshold functions σ_c, which are satisfied by many choices of σ_c in practice. A prominent example is the logistic (or sigmoid) function in which case $\sigma_c(z) = 1/(1 + e^{-z})$, which is also used in deep learning. Another possible choice is the *probit function*, which corresponds to $\sigma_c(t) = \Phi(t)$, where Φ is the cumulative distribution function of the standard normal distribution.

Assumption 2. *The threshold function* σ_c *is twice continuously differentiable on* $\mathbb{R}$. *It is bounded and has bounded first and second derivatives; that is, there exists a constant* $L_{\sigma_c} > 0$ *such that* $\max\{\|\sigma_c\|_\infty, \|\sigma_c'\|_\infty, \|\sigma_c''\|_\infty\} \le L_{\sigma_c}$. *The distribution of the input data* A_{in} *has compact support; that is,* $\|A_{in}\| \le D$ *for some* $D > 0$.

We show in the next lemma that, if Assumption 2 holds, then Assumption 1 holds with explicit constants A_0, B, M, m, b, and σ_c that are precise. The proof is in the e-companion.

Lemma 1. *In the setting of binary linear classification, consider the SGHMC method applied to* Objective (41), *where gradients are estimated according to* (42), *where the probability law* μ_0 *of the initial state has compact support. If Assumption 2 holds, then Assumption 1 holds for any* $\delta \in [\frac{1}{4n_b}, 1)$ *with the following constants:*

$$A_0 = (1 + \|\sigma_c(0)\|)^2, \quad B = 2D(1 + \|\sigma_c\|_\infty)\|\sigma_c'\|_\infty, \tag{43}$$

$$M = 2D^2\|\sigma_c'\|_\infty^2 + 2D^2(1 + \|\sigma_c\|_\infty)\|\sigma_c''\|_\infty + 5\lambda_r, \tag{44}$$

$$m = \lambda/2, \quad b = 8(1 + \|\sigma_c\|_\infty)^2\|\sigma_c'\|_\infty^2 D^2/\lambda_r. \tag{45}$$

We conclude from Lemma 1 that the objective is dissipative, and our main results for the SGHMC1 and SGHMC2 algorithms described in Sections 3–5 apply to binary linear classification under Assumption 2 with the constants given in Lemma 1 and in which μ_* is given by the formula in Table 1. For example, if $D = \mathcal{O}(1)$, then we have $\frac{1}{\mu_*} = \tilde{\Theta}(\sqrt{d+\beta}e^{\tilde{\Theta}(d+\beta)})$ (see (EC.89)), and we conclude from (39) that it suffices to have

$$\hat{K}_{SGHMC2} = \tilde{\Omega}\left(\frac{d^9}{\mu_*^4\hat{\varepsilon}^6}\right) = \tilde{\Omega}\left(\frac{d^9 e^{\tilde{\Theta}(d+\beta)}}{(d^2+\beta^2)\hat{\varepsilon}^6}\right)$$

stochastic gradient evaluations to converge to an $\hat{\varepsilon}$ neighborhood of an almost ERM ignoring the log factors in the parameters $\hat{\varepsilon}, \mu_*, d$ and hiding other constants that can be made explicit based on Lemma 1.

We also note that, under further assumptions on the statistical nature of the input and if the number of data points is large enough, it can be shown that Objective (41) admits a unique minimizer, the objective is strongly convex in some regions (Mei et al. 2018), and the convergence to the unique minimizer is independent of the dimension d. However, our assumptions here are weaker; for example, we have weaker assumptions on the threshold function σ_c. Therefore, such arguments are not directly applicable.

6.2. Robust Ridge Regression

Given an input (feature) vector $A_{in} \in \mathbb{R}^d$, the aim is to predict the output $Y \in \mathbb{R}$. Given access to a data set of input–output pairs $z_i = (a_i, y_i)$, we assume a linear model $y_i = a_i^T\tilde{x} + \varepsilon_i$, where the errors ε_i are i.i.d. with mean zero. The standard *ridge regression* estimate of $\tilde{x}$ minimizes a penalized residual sum of squares (Hoerl and Kennard 1970); that is, it minimizes $\sum_{i=1}^n \|y_i - \langle x, a_i\rangle\|^2 + \lambda_r\|x\|^2$, where $\lambda_r > 0$ is a regularization parameter.[5] However, this formulation can be sensitive to outliers. Robust formulations of the ridge regression (Razavi et al. 2012) can be obtained if one solves instead the following problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f(x, z_i), \quad f(x, z_i) = \rho(y_i - \langle x, a_i\rangle) + \frac{\lambda_r}{2}\|x\|^2, \tag{46}$$

where $\lambda_r > 0$ is a regularization parameter and $\rho : \mathbb{R} \to \mathbb{R}$ is a suitably chosen loss function. In particular, for achieving robustness to outliers, the nonconvex choices of the function ρ that are either bounded or slowly growing near infinity has been considered in the literature (as opposed to the standard ridge regression setting that corresponds to $\rho(t) = \|t\|^2$). For example, popular choices of the function $t \mapsto \rho(t)$ include *Tukey's bisquare loss* defined as

$$\rho_{\text{Tukey}}(t) = \begin{cases} 1 - (1 - (t/t_0)^2)^3 & \text{for } \|t\| \le t_0, \\ 1 & \text{for } |t| \ge t_0, \end{cases}$$

(see, e.g., Mei et al. 2018) and exponential squared loss (Wang et al. 2013): $\rho_{\exp}(t) = 1 - e^{-\|t\|^2/t_0}$, where $t_0 > 0$ is a tuning parameter. In the following, similar to Wang et al. (2019), we assume that the data A_{in} is bounded and the threshold function and its derivatives up to order two are bounded. This assumption for ρ is satisfied in several cases, including Tukey's bisquare and exponential squares losses.

Assumption 3. *The function ρ is twice continuously differentiable on $\mathbb{R}$. The function ρ is bounded and has bounded first and second derivatives; that is, there exists a constant L_ρ such that $\max(\|\rho\|_\infty, \|\rho'\|_\infty, \|\rho''\|_\infty) \le L_\rho$. Furthermore, the distribution of the input data A_{in} has compact support; that is, there exists D such that $\|A_{in}\| \le D$.*

The following lemma shows that, under Assumption 3, our assumptions (Assumption 1) for analyzing the SGHMC methods hold with proper initialization.

Lemma 2. *In the setting of robust regression, consider* Objective (41) *in which gradients are estimated according to* (42) *in which the probability law μ_0 of the initial state has compact support. If Assumption 2 holds, then Assumption 1 holds for both SGHMC1 and SGHMC2 methods for any choice of $\delta \in [\frac{1}{4n_b}, 1)$ with the following constants:*

$$A_0 = \|\rho\|_\infty, \qquad B = 4\|\rho'\|_\infty D, \tag{47}$$

$$M = \|\rho''\|_\infty D^2 + \lambda_r, \qquad m = \lambda_r/2, \qquad b = \frac{2\|\rho'\|_\infty^2 D^2}{\lambda_r}. \tag{48}$$

Similarly, we conclude from Lemma 2 that our main results for the SGHMC1 and SGHMC2 algorithms described in Sections 3–5 apply to the problem of robust regression under Assumption 3.

7. Outline of the Proof

To obtain the main results in this paper, we adapt the proof techniques of Raginsky et al. (2017) developed for the overdamped dynamics to the underdamped dynamics and combine it with the analysis of Eberle et al. (2019), which quantifies the convergence rate of the underdamped Langevin SDE to its equilibrium. In an analogy to the fact that momentum-based first order optimization methods require a different Lyapunov function and a quite different set of analysis tools (compared with their nonaccelerated variants) to achieve fast rates (see, e.g., Nesterov 1983, Su et al. 2014, Lu et al. 2018), our analysis of the momentum-based SGHMC1 and SGHMC2 algorithms requires studying a different Lyapunov function $\mathcal{V}$ defined in (13) that also depends on the objective f as opposed to the classic Lyapunov function $\mathcal{H}(x) = \|x\|^2$ arising in the study of the SGLD algorithm (see, e.g., Mattingly et al. 2002, Raginsky et al. 2017). This fact introduces some challenges for the adaptation of the existing

analysis techniques for SGLD to SGHMC. For this purpose, we take the following steps.

First, we show that SGHMC1 and SGHMC2 iterates track the underdamped Langevin diffusion closely in the two-Wasserstein metric. As this metric requires finiteness of second moments, we first establish uniform (in time) L^2 bounds for both the underdamped Langevin SDE and SGHMC1 and SGHMC2 iterates (see Lemmas EC.1 and EC.5), exploiting the structure of the Lyapunov function $\mathcal{V}$. Second, we obtain a bound for the Kullback–Leibler divergence between the discrete and continuous underdamped dynamics making use of the Girsanov theorem, which is then converted to bounds in the two-Wasserstein metric by an application of an optimal transportation inequality of Bolley and Villani (2005). This step requires proving a certain exponential integrability property of the underdamped Langevin diffusion (Lemma EC.2). We show in Lemma EC.2 that the exponential moments grow at most linearly in time, which strictly improves the exponential growth in time in lemma 4 in Raginsky et al. (2017).[6] As a result, the method improves upon the ε dependence of the number of iterates (see Equations (32) and (33)).

Second, we apply the seminal result of Eberle et al. (2019), which shows that the continuous-time underdamped Langevin SDE is geometrically ergodic with an explicit rate μ_* in the two-Wasserstein metric. In order to get explicit performance guarantees, we derive new bounds that make the dependence of the constants to the initialization in Eberle et al. (2019) explicit (see Lemma EC.4).

As the x-marginal of the equilibrium distribution $\pi_{\mathbf{z}}(dx,dv)$ of the underdamped Langevin SDE concentrates around the global minimizers of $F_{\mathbf{z}}$ for β appropriately chosen and we can control the error between the discrete-time SGHMC1 and SGHMC2 dynamics and the underdamped SDE by choosing the step size accordingly, this leads to performance bounds for the empirical risk minimizations for the SGHMC1 and SGHMC2 algorithms in Corollaries 1 and 3. For controlling the population risk during SGHMC iterations, in addition to the empirical risk, one has to control the *generalization error* $F(X_k) - F_{\mathbf{Z}}(X_k)$ that accounts for the differences between the finite sample size Problem (2) and the original Problem (1). By exploiting the fact that the $x-$ marginal of the invariant distribution for the underdamped dynamics is the same as it is in the overdamped case, we control the generalization error in Corollaries 2 and 4, which is no worse than that of the available bounds for SGLD given in Raginsky et al. (2017).

8. Conclusion

SGHMC is a momentum-based popular variant of the stochastic gradient in which a controlled amount of isotropic Gaussian noise is added to the gradient estimates for optimizing a nonconvex function. We obtain first-time, finite-time guarantees for the convergence of SGHMC1 and SGHMC2 algorithms to the ε-global minimizers under some regularity assumption on the nonconvex objective f. We also show that, on a class of nonconvex problems, SGHMC2 can be faster than overdamped Langevin MCMC approaches, such as SGLD, in the sense that the best available bounds for SGHMC2, which we prove in our paper, are better than the best available bounds for SGLD. This effect is due to the momentum term in the underdamped SDE. Furthermore, our results show that momentum-based acceleration is possible on a class of nonconvex problems under some conditions if we compare known upper bounds between SGLD and SGHMC. Finally, we mention a few limitations in our work that may lead to some future research directions. In our paper, the performance dependence on dimension is exponential in general. In the future, we will investigate for what class of (nonconvex) target functions f we can obtain performance bounds independent of dimension d or has polynomial dependence on d. In addition, our results suggest that momentum-based SGHMC methods work particularly well when the (nonconvex) target functions have relatively flat landscapes. In the future, we will investigate whether we can obtain theoretical results for SGHMC on a wider class of nonconvex problems.

Acknowledgments

The authors thank the area editor, the associate editor, and an anonymous referee for helpful comments and suggestions. They also thank Agostino Capponi, Xiuli Chao, Wenbin Chen, Jim Dai, Murat A. Erdogdu, Qi Feng, Fuqing Gao, Jianqiang Hu, Jin Ma, Sanjoy Mitter, Asuman Ozdaglar, Pablo Parrilo, Umut Şimşekli, and S. R. S. Varadhan for helpful discussions. The authors are listed in alphabetical order.

Endnotes

[1] We note that, in our notation, $\mathbf{Z}$ is a random vector, whereas $\mathbf{z}$ is a deterministic vector associated to a data set that corresponds to a realization of the random vector $\mathbf{Z}$.

[2] With slight abuse of notation, we use $\pi_{\mathbf{z}}(dx)$ to denote the x-marginal of the equilibrium distribution $\pi_{\mathbf{z}}(dx,dv)$.

[3] In Raginsky et al. (2017), their formula for λ_* missed the β^{-1} factor.

[4] We emphasize that the effect of the last term $\sqrt{\log\log(1/\varepsilon)}$ appearing in (31) is typically negligible compared with other parameters. For instance, even if $\varepsilon = 2^{-2^{16}}$ is double-exponentially small, we have $\sqrt{\log\log(1/\varepsilon)} \le 4$.

[5] See Kilmer and O'Leary (2001) for details regarding the choice of the parameter λ_r.

[6] The method used in the proof of Lemma EC.2 can indeed be adapted to improve the exponential integrability and, hence, the overall estimates in Raginsky et al. (2017) for SGLD as well.

References

Ahn S, Korattikara A, Welling M (2012) Bayesian posterior sampling via stochastic gradient Fisher scoring. *Internat. Conf. Machine Learn.*, 1771–1778.

Allen-Zhu Z, Hazan E (2016) Variance reduction for faster non-convex optimization. Balcan MF, Weinberger KQ, eds. *Proc. 33rd Internat. Conf. Machine Learn.*, vol. 48 (PMLR, New York), 699–707.

Anil C, Lucas J, Grosse R (2019) Sorting out Lipschitz function approximation. Chaudhuri K, Salakhutdinov R, eds. *Proc. 36th Internat. Conf. Machine Learn.*, vol. 97 (PMLR, New York), 291–301.

Belloni A, Liang T, Narayanan H, Rakhlin A (2015) Escaping the local minima via simulated annealing: Optimization of approximately convex functions. *Conf. Learn. Theory*, 240–265.

Bertsimas D, Tsitsiklis J (1993) Simulated annealing. *Statist. Sci.* 8(1): 10–15.

Betancourt M (2017) A conceptual introduction to Hamiltonian Monte Carlo. Preprint, submitted January 10, https://arxiv.org/abs/1701.02434.

Betancourt M, Byrne S, Girolami M (2014) Optimizing the integrator step size for Hamiltonian Monte Carlo. Preprint, submitted November 24, https://arxiv.org/abs/1411.6669.

Betancourt M, Byrne S, Livingstone S, Girolami M (2017) The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* 23(4A):2257–2298.

Bolley F, Villani C (2005) Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des Sciences de Toulouse: Mathématiques* 14(3):331–352.

Borkar VS, Mitter SK (1999) A strong approximation theorem for stochastic recursive algorithms. *J. Optim. Theory Appl.* 100(3): 499–513.

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

Carmon Y, Duchi J, Hinder O, Sidford A (2018) Accelerated methods for nonconvex optimization. *SIAM J. Optim.* 28(2):1751–1772.

Chapelle O, Do CB, Teo CH, Le QV, Smola AJ (2009) Tighter bounds for structured estimation. Koller D, Schuurmans D, Bengio Y, Bottou L, eds. *Advances in Neural Information Processing Systems*, vol. 21 (Curran Associates, Inc.), 281–288.

Chatterji NS, Flammarion N, Ma YA, Bartlett PL, Jordan MI (2018) On the theory of variance reduction for stochastic gradient Monte Carlo. *Internat. Conf. Machine Learn.*, 764–773.

Chaudhari P, Choromanska A, Soatto S, LeCun Y, Baldassi C, Borgs C, Chayes J, Sagun L, Zecchina R (2017) Entropy-SGD: Biasing gradient descent into wide valleys. *Internat. Conf. Learn. Representations*.

Chen C, Ding N, Carin L (2015) On the convergence of stochastic gradient MCMC algorithms with high-order integrators. *Proc. 28th Internat. Conf. Neural Inform. Processing Systems*, vol. 2, 2278–2286.

Chen T, Fox E, Guestrin C (2014) Stochastic gradient Hamiltonian Monte Carlo. Xing EP, Jebara T, eds. *Proc. 31st Internat. Conf. Machine Learn.*, vol. 32 (PMLR, Beijing), 1683–1691.

Chen C, Carlson D, Gan Z, Li C, Carin L (2016) Bridging the gap between stochastic gradient MCMC and stochastic optimization. *Proc. 19th Internat. Conf. Artificial Intelligence Statist.*, 1051–1060.

Cheng X, Chatterji NS, Bartlett PL, Jordan MI (2018a) Underdamped Langevin MCMC: A non-asymptotic analysis. *Proc. 31st Annual Conf. Learn. Theory*.

Cheng X, Chatterji NS, Abbasi-Yadkori Y, Bartlett PL, Jordan MI (2018b) Sharp convergence rates for Langevin dynamics in the nonconvex setting. Preprint, submitted May 4, https://arxiv.org/abs/1805.01648.

Chiang TS, Hwang CR, Sheu SJ (1987) Diffusion for global optimization in $\mathbb{R}^n$. *SIAM J. Control Optim.* 25(3):737–753.

Collobert R, Sinz F, Weston J, Bottou L (2006) Trading convexity for scalability. *Proc. 23rd Internat. Conf. Machine Learn.*, 201–208.

Dalalyan AS (2017) Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. Roy. Statist. Soc. Ser. B* 79(3):651–676.

Dalalyan AS, Karagulyan AG (2019) User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes Their Appl.* 129(12):5278–5311.

Dalalyan AS, Riou-Durand L (2020) On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli* 26(3):1956–1988.

Dalalyan AS, Karagulyan A, Riou-Durand L (2019) Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. Preprint, submitted June 20, https://arxiv.org/abs/1906.08530.

Ding N, Fang Y, Babbush R, Chen C, Skeel RD, Neven H (2014) Bayesian sampling using stochastic gradient thermostats. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc.), 3203–3211.

Du SS, Lee JD, Tian Y, Singh A, Poczos B (2018) Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. *Internat. Conf. Machine Learn.*, 1339–1348.

Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. *Phys. Lett. B.* 195(2):216–222.

Eberle A, Guillin A, Zimmer R (2019) Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.* 47(4): 1982–2010.

Foster DJ, Sekhari A, Sridharan K (2018) Uniform convergence of gradients for non-convex learning and optimization. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc.) 31:8745–8756.

Ge R, Lee JD, Ma T (2018) Learning one-hidden-layer neural networks with landscape design. *Internat. Conf. Learn. Representations.*

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

Ghadimi S, Lan G (2016) Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Programming* 156(1):59–99.

Gidas B (1985) Nonstationary Markov chains and convergence of the annealing algorithm. *J. Statist. Phys.* 39(1–2):73–131.

Hajek B (1985) A tutorial survey of theory and applications of simulated annealing. *24th IEEE Conf. Decision Control* (IEEE), 755–760.

Hale J (1988) *Asymptotic Behavior of Dissipative Systems*, vol. 25 (American Mathematical Society).

Hardt M, Recht B, Singer Y (2016) Train faster, generalize better: stability of stochastic gradient descent. *Internat. Conf. Machine Learn.*, 1225–1234.

Hazan E, Levy KY, Shalev-Shwartz S (2016) On graduated optimization for stochastic non-convex problems. *Internat. Conf. Machine Learn.*, 1833–1841.

Hérau F, Nier F (2004) Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Arch. Rational Mechanics Anal.* 171(2):151–218.

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.

Holley RA, Kusuoka S, Stroock DW (1989) Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Functional Anal.* 83(2):333–347.

Hu Y, Li C, Meng K, Qin J, Yang X (2017) Group sparse optimization via $\ell_{p,q}$ regularization. *J. Machine Learn. Res.* 18(1):960–1011.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning*, vol. 112 (Springer).

Jofré A, Thompson P (2019) On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Math. Programming* 174:253–292.

Kilmer ME, O'Leary DP (2001) Choosing regularization parameters in iterative methods for ill-posed problems. *SIAM J. Matrix Anal. Appl.* 22(4):1204–1221.

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Sci.* 220(4598):671–680.

Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7(4):284–304.

Lee YT, Vempala SS (2018) Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. *Proc. 50th Annual ACM SIGACT Sympos. Theory Comput.* (ACM), 1115–1121.

Leimkuhler B, Matthews C, Stoltz G (2015) The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA J. Numerical Anal.* 36(1):13–79.

Liu T, Chen Z, Zhou E, Zhao T (2018) Toward deeper understanding of nonconvex stochastic optimization with momentum using diffusion approximations. Preprint, submitted February 14, https://arxiv.org/abs/1802.05155.

Lu H, Freund RM, Mirrokni V (2018) Accelerating greedy coordinate descent methods. *Internat. Conf. Machine Learn.*, 3257–3266.

Ma YA, Chen T, Fox E. (2015) A complete recipe for stochastic gradient MCMC. Cortes C, Lawrence N, Lee D, Sugiyama M, GarnettR, eds. *Advances in Neural Information Processing Systems*, vol. 28 (Curran Associates, Inc.), 2917–2925.

Mangoubi O, Smith A (2017) Rapid Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. Preprint, submitted August 23, https://arxiv.org/abs/1708.07114.

Mangoubi O, Pillai NS, Smith A (2018) Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? Preprint, submitted August 9, https://arxiv.org/abs/1808.03230.

Mattingly JC, Stuart AM, Higham DJ (2002) Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise. *Stochastic Processes Their Appl.* 101(2):185–232.

Mei S, Bai Y, Montanari A (2018) The landscape of empirical risk for nonconvex losses. *Ann. Statist.* 46(6A):2747–2774.

Neal R (2010) MCMC using Hamiltonian dynamics. Brooks S, Gelman A, Jones G, Meng X-L, eds. *Handbook of Markov Chain Monte Carlo* (Chapman & Hall / CRC press, Boca Raton, Florida), 113–162.

Nesterov YE (1983) A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR.* 269:543–547.

Nguyen T, Sanner S (2013) Algorithms for direct 0–1 loss optimization in binary classification. *Internat. Conf. Machine Learn.*, 1085–1093.

O'Neill M, Wright SJ (2019) Behavior of accelerated gradient methods near critical points of nonconvex problems. *Math. Programming* 176:403–427.

Patterson S, Teh YW (2013) Stochastic gradient Riemannian Langevin dynamics on the probability simplex. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates, Inc.), 3102–3110.

Pavliotis GA (2014) *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, vol. 60 (Springer).

Polyak BT (1987) *Introduction to Optimization* (Optimization Software, New York).

Raginsky M, Rakhlin A, Telgarsky M (2017) Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. *Conf. Learn. Theory*, 1674–1703.

Razavi SA, Ollila E, Koivunen V (2012) Robust greedy algorithms for compressed sensing. *Proc. 20th Eur. Signal Processing Conf.* (IEEE), 969–973.

Shi B, Du SS, Jordan MI, Su WJ (2018) Understanding the acceleration phenomenon via high-resolution differential equations. Preprint, submitted October 21, https://arxiv.org/abs/1810.08907.

Şimşekli U, Badeau R, Cemgil T, Richard G (2016) Stochastic quasi-Newton Langevin Monte Carlo. *Internat. Conf. Machine Learn.*, 642–651.

Şimşekli U, Yildiz Ç, Nguyen TH, Cemgil T, Richard G (2018) Asynchronous stochastic quasi-Newton MCMC for non-convex optimization. Dy J, Krause A, eds. *Proc. 35th Internat. Conf. Machine Learn.*, vol. 80 (PMLR), 4674–4683.

Su W, Boyd S, Candes E (2014) A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc.), 2510–2518.

Sun J, Qu Q, Wright J (2016) Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Trans. Inform. Theory* 63(2):853–884.

Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. Dasgupta S, McAllester D, eds. *Proc. 30th Internat. Conf. Machine Learn.*, vol. 28 (PMLR, Atlanta), 1139–1147.

Teh YW, Thiery AH, Vollmer SJ (2016) Consistency and fluctuations for stochastic gradient Langevin dynamics. *J. Machine Learn. Res.* 17(1):193–225.

Tzen B, Liang T, Raginsky M (2018) Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. *Conf. Learn. Theory*, 857–875.

Villani C (2008) *Optimal Transport: Old and New*, vol. 338 (Springer Science & Business Media, Springer-Verlag Berlin Heidelberg).

Wang D, Chen C, Xu J (2019) Differentially private empirical risk minimization with non-convex loss functions. Chaudhuri K, Salakhutdinov R, eds. *Proc. 36th Internat. Conf. Machine Learn.*, vol 97 (PMLR), 6526–6535.

Wang X, Jiang Y, Huang M, Zhang H (2013) Robust variable selection with exponential squared loss. *J. Amer. Statist. Assoc.* 108(502): 632–643.

Welling M, Teh YW (2011) Bayesian learning via stochastic gradient Langevin dynamics. *Proc. 28th Internat. Conf. Machine Learn.* (Citeseer), 681–688.

Wibisono A (2018) Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. Bubeck S, Perchet V, Rigollet P, eds. *Proc. 31st Annual Conf. Learn. Theory*, vol. 75 (PMLR), 2093–3027.

Wilson AC, Recht B, Jordan MI (2016) A Lyapunov analysis of momentum methods in optimization. Preprint, submitted November 8, https://arxiv.org/abs/1611.02635.

Wu Y, Liu Y (2007) Robust truncated hinge loss support vector machines. *J. Amer. Statist. Assoc.* 102(479):974–983.

Xu P, Chen J, Zou D, Gu Q (2018) Global convergence of Langevin dynamics based algorithms for nonconvex optimization. Bengio S, Wallach H, Larochelle H, Grauman K. Cesa-Bianchi N, Garnett R, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc.) 31:3122–3133.

Zhang Y, Liang P, Charikar M (2017a) A hitting time analysis of stochastic gradient Langevin dynamics. Kale S, Shamir O, eds. *Proc. 2017 Conf. Learning Theory*, vol. 65 (PMLR), 1980–2022.

Zhang H, Zhou Y, Liang Y, Chi Y (2017b) A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *J. Machine Learn. Res.* 18(1):5164–5198.

Xuefeng Gao is an associate professor at the department of systems engineering and engineering management at The Chinese University of Hong Kong.

Mert Gürbüzbalaban is an assistant professor at the department of management and information systems of the Rutgers Business School.

Lingjiong Zhu is an associate professor at the Department of Mathematics, Florida State University.