Journal of Machine Learning Research 24 (2023) 1-53 Submitted 4/20; Revised 12/23; Published 12/23

On Learning Rates and Schrödinger Operators

Bin Shi SHIBIN@LSEC.CC.AC.CN
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190, China

School of Mathematical Sciences
University of Chinese Academy of Sciences
Beijing, 100049, China

Weijie J. Su SUW@WHARTON.UPENN.EDU
Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: Zaid Harchaoui

Abstract

Understanding the iterative behavior of stochastic optimization algorithms for minimizing nonconvex functions remains a crucial challenge in demystifying deep learning. In particular, it is not yet understood why certain simple techniques are remarkably effective for tuning the learning rate in stochastic gradient descent (SGD), arguably the most basic optimizer for training deep neural networks. This class of techniques includes learning rate decay, which begins with a large initial learning rate and is gradually reduced. In this paper, we present a general theoretical analysis of the effect of the learning rate in SGD. Our analysis is based on the use of a *learning-rate-dependent stochastic differential equation* (LR-dependent SDE) as a tool that allows us to set SGD distinctively apart from both gradient descent and stochastic gradient Langevin dynamics (SGLD). In contrast to prior research, our analysis builds on the analysis of a partial differential equation that models the evolution of probability densities, drawing insights from Wainwright and Jordan (2006); Jordan (2018). From this perspective, we derive the linear convergence rate of the probability densities, highlighting its dependence on the learning rate. Moreover, we obtain an explicit expression for the optimal linear rate by analyzing the spectrum of the Witten-Laplacian, a special case of the Schrödinger operator associated with the LR-dependent SDE. This expression clearly reveals the dependence of the linear convergence rate on the learning rate—the linear rate decreases rapidly to zero as the learning rate tends to zero for a broad class of nonconvex functions, whereas it stays constant for strongly convex functions. Based on this sharp distinction between nonconvex and convex problems, we provide a mathematical interpretation of the benefits of using learning rate decay for nonconvex optimization.

©2023 Bin Shi and Weijie J. Su and Michael I. Jordan.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/20-364.html.

Keywords: Deep Neural Networks, Nonconvex Optimization, Stochastic Gradient Descent, LR-dependent SDE, Fokker–Planck–Smoluchowski equation, Schrödinger Operator, Witten-Laplacian, Learning Rate.

1. Introduction

Gradient-based optimization has been the workhorse algorithm powering recent developments in statistical machine learning. Many of these developments involve solving nonconvex optimization problems, which raises new challenges for theoreticians, given that classical theory has often been restricted to the convex setting.

A particular focus in machine learning is the class of gradient-based methods referred to as *stochastic gradient descent* (SGD), given its desirable runtime properties, and its desirable statistical performance in a wide range of nonconvex problems. Consider the minimization of a (nonconvex) function f defined in terms of an expectation:

$$f(x) = \mathbb{E}_\zeta f(x;\zeta),$$

where the expectation is over the randomness embodied in ζ. A simple example is empirical risk minimization, where the loss function,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

is averaged over n data points, where the datapoint-specific losses, $f_i(x)$, are indexed by i and where x denotes a parameter. When n is large, it is computationally prohibitive to obtain the full gradient of the objective function, and SGD provides a compelling alternative. SGD is a gradient-based update based on a (noisy) gradient evaluated from a single data point or a mini-batch:

$$\widetilde{\nabla} f(x) := \frac{1}{B}\sum_{i\in\mathcal{B}} \nabla f_i(x) = \nabla f(x) + \xi,$$

where the set $\mathcal{B}$ of size B is sampled uniformly from the n data points and therefore the noise term ξ has mean zero. Starting from an initial point x_0, SGD updates the iterates according to

$$x_{k+1} = x_k - s\widetilde{\nabla} f(x_k) = x_k - s\nabla f(x_k) - s\xi_k, \tag{1}$$

where ξ_k denotes the noise term at the kth iteration. Note that the step size $s > 0$, also known as the *learning rate*, can either be constant or vary with the iteration Bottou (2010).

The learning rate plays an essential role in determining the performance of SGD and many of the practical variants of SGD Bengio (2012).[1] The overall effect of the learning rate can be complex. In convex optimization problems, theoretical analysis can explain many aspects of this complexity, but in the nonconvex setting the effect of the learning rate is yet more complex and theory is lacking Zeiler (2012); Kingma and Ba (2014). As a numerical illustration of this complexity, Figure 1 plots the error of SGD with a piecewise constant learning rate in the training of a neural network on the CIFAR-10 dataset. With

1. Note that the mini-batch size as another parameter can be, to some extent, incorporated into the learning rate. See the discussion later in this section.

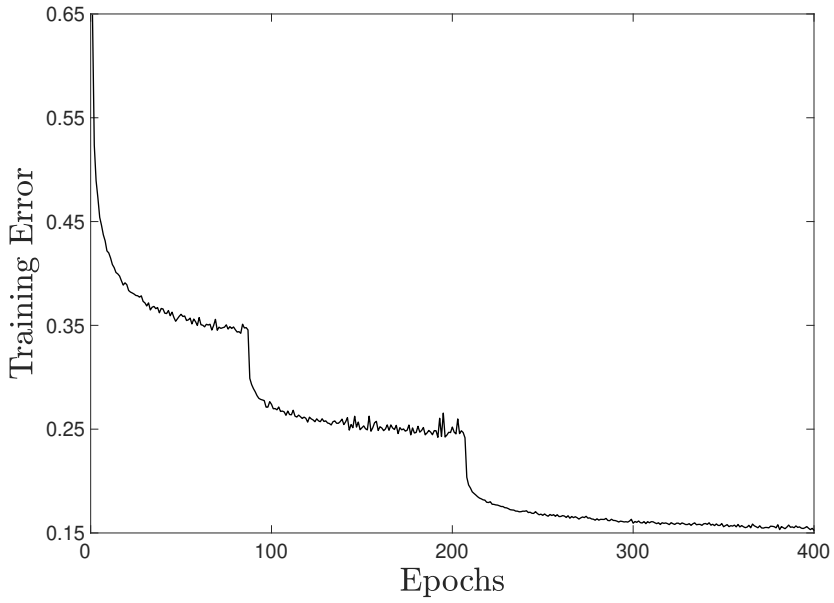

Figure 1: Training error using SGD with mini-batch size 32 to train an 8-layer convolutional neural network on CIFAR-10 Krizhevsky (2009). The first 90 epochs use a learning rate of $s = 0.006$, the next 120 epochs use $s = 0.003$, and the final 190 epochs use $s = 0.0005$. Note that the training error decreases as the learning rate s decreases and a smaller s leads to a larger number of epochs for SGD to reach a plateau. See He et al. (2016) for further investigation of this phenomenon.

a constant learning rate, SGD quickly reaches a plateau in terms of training error, and whenever the learning rate decreases, the plateau decreases as well, thereby yielding better optimization performance. This illustration exemplifies the idea of learning rate decay, a technique that is used in training deep neural networks (see, e.g., He et al., 2016; Bottou et al., 2018; Sordello and Su, 2019). Despite its popularity and the empirical evidence of its success, however, the literature stops short of providing a *general* and *quantitative* approach to understanding how the learning rate impacts the performance of SGD and its variants in the nonconvex setting You et al. (2019); Li et al. (2019b). Accordingly, strategies for setting learning rate decay schedules are generally ad hoc and empirical.

In the current paper, we provide theoretical insight into the dependence of SGD on the learning rate in nonconvex optimization. Our approach builds on a recent line of work in which optimization algorithms are studied via the analysis of their behavior in continuous-time limits Su et al. (2016); Jordan (2018); Shi et al. (2018). Specifically, in the case of SGD, we study stochastic differential equations (SDEs) as surrogates for discrete stochastic optimization methods (see, e.g., Kushner and Yin, 2003; Li et al., 2017; Krichene and Bartlett, 2017; Chaudhari et al., 2018; Diakonikolas and Jordan, 2019). The construction is roughly as follows. Taking a small but nonzero learning rate s, let $t_k = ks$ denote a time step and define $x_k = X_s(t_k)$ for some sufficiently smooth curve $X_s(t)$. Applying a Taylor

expansion in powers of s, we obtain:

$$x_{k+1} = X_s(t_{k+1}) = X_s(t_k) + \dot{X}_s(t_k)s + O(s^2).$$

Let W be a standard Brownian motion, where we assume that the noise term ξ_k is approximately normally distributed with unit variance. Informally, this leads to[2]

$$-\sqrt{s}\xi_k = W(t_{k+1}) - W(t_k) = s\,\frac{\mathrm{d}W(t_k)}{\mathrm{d}t} + O(s^2).$$

Plugging the last two displays into (1), we get

$$\dot{X}_s(t_k) + O(s) = -\nabla f(X_s(t_k)) + \sqrt{s}\frac{\mathrm{d}W(t_k)}{\mathrm{d}t} + O\left(s^{\frac{3}{2}}\right).$$

Retaining both $O(1)$ and $O(\sqrt{s})$ terms but ignoring smaller terms, we obtain a *learning-rate-dependent stochastic differential equation* (LR-dependent SDE) that approximates the discrete-time SGD algorithm:

$$\mathrm{d}X_s = -\nabla f(X_s)\mathrm{d}t + \sqrt{s}\mathrm{d}W, \tag{2}$$

where the initial condition is the same value x_0 as its discrete counterpart. More generally, Li et al. (2019a); Chaudhari and Soatto (2018) consider SDEs with variable-dependent noise covariance as approximating surrogates for SGD. The LR-dependent SDE (2) is a convenient simple model that allows for a fine-grained analysis, as we will show in this paper. As an indication of the generality of this formulation, we note that it can seamlessly take account of the mini-batch size B; in particular, the effective learning rate scales as $O(s/B)$ in the mini-batch setting (see more discussion in Smith et al. (2017)). Throughout this paper we focus on (2) and regard s alone as the effective learning rate.[3]

Intuitively, a larger learning rate s gives rise to more stochasticity in the LR-dependent SDE (2), and vice versa. Accordingly, the learning rate must have a substantial impact on the dynamics of SGD in its continuous-time formulation. In stark contrast, this parameter plays a fundamentally different role on gradient descent (GD) and stochastic gradient Langevin dynamics (SGLD) when one considers their approximating differential equations. In particular, consider GD:

$$x_{k+1} = x_k - s\nabla f(x_k),$$

which can be modeled by the following ordinary differential equation (ODE):

$$\dot{X} = -\nabla f(X),$$

and the SGLD algorithm, which adds Gaussian noise ξ_k to the GD iterates:

$$x_{k+1} = x_k - s\nabla f(x_k) + \sqrt{s}\xi_k,$$

2. Although a Brownian motion is not differentiable, the formal notation $\mathrm{d}W(t)/\mathrm{d}t$ can be given a rigorous interpretation Evans (2012); Villani (2006).
3. Recognizing that the variance of ξ_k is inversely proportional to the mini-batch size B, we assume that the noise term ξ_k has variance σ^2/B. Under this assumption the resulting SDE reads $\mathrm{d}X_s = -\nabla f(X_s)\mathrm{d}t + \sigma\sqrt{s/B}\mathrm{d}W$. In light of this, the effective learning rate through incorporating the mini-batch size is $O(\sigma^2 s/B)$.

and its SDE model:

$$\mathrm{d}X = -\nabla f(X)\mathrm{d}t + \mathrm{d}W.$$

These differential equations are derived in the same way as (2), namely by the Taylor expansion and retaining $O(1)$ and $O(\sqrt{s})$ terms.[4] While the SDE for modeling SGD sets the square root of the learning rate to be its diffusion coefficient, both the GD and SGLD counterparts are completely free of this parameter. This distinction between SGD and the other two methods is reflected in their different numerical performance as revealed in Figure 2. The right plot of this figure shows that the behaviors of both GD and SGLD in the time $t = ks$ scale are almost invariant in terms of optimization error with respect to the learning rate. In striking contrast, the stationary optimization error of SGD decreases significantly as the learning rate decays. As a consequence of this distinction, GD and SGLD do not exhibit the phenomenon that is shown in Figure 1.

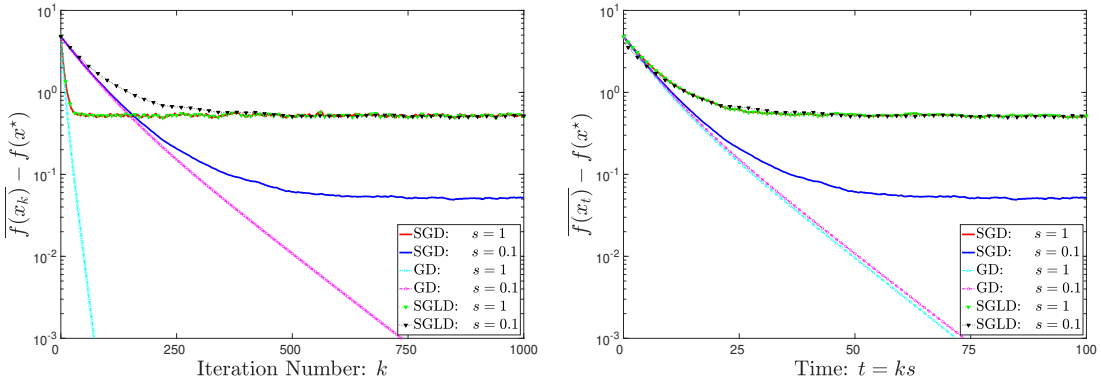

Figure 2: Illustrative examples showing distinct behaviors of GD, SGD, and SGLD. The y-axis displays the optimization error $\overline{f(x_k)} - f(x^\star)$, where $f(x^\star)$ denotes the minimum value of the objective and in the case of SGD and SGLD $\overline{f(x_k)}$ denotes an average over 1000 replications. The objective function is $f(x_1, x_2) = 5 \times 10^{-2}x_1^2 + 2.5 \times 10^{-2}x_2^2$, with an initial point $(8, 8)$, and the noise ξ_k in the gradient follows a standard normal distribution. Note that SGD with $s = 1$ is identical to SGLD with $s = 1$. As shown in the right panel, taking time $t = ks$ as the x-axis, the learning rate has little to no impact on GD and SGLD in terms of optimization error.

1.1 Overview of contributions

The discussion thus far suggests that one may examine the effect of the learning rate in SGD using the LR-dependent SDE (2). In particular, this SDE distinguishes SGD from GD and SGLD. Accordingly, in the current paper, we study the LR-dependent SDE and make the following contributions.

1. **LR-dependent Fokker–Planck–Smoluchowski equation.** The perspective of considering the evolution behavior of probability distributions over points instead of

4. The coefficients of the $O(\sqrt{s})$ terms turn out to be zero in both differential equations. See more discussion in Appendix A.1 and particularly Figure 12 therein.

a single point is proposed in Wainwright and Jordan (2006); Jordan (2018). In this paper, we instantiate this perspective for SGD via the LR-dependent SDE and use existing techniques to derive the governing LR-dependent Fokker–Planck–Smoluchowski equation for the evolution of the probability densities. By utilizing the error decomposition in Raginsky et al. (2017), we show that, for a large class of (nonconvex) objectives, the continuous-time formulation of SGD converges to its stationary distribution at a *linear rate*.[5] In particular, the solution $X_s(t)$ to the LR-dependent SDE obeys

$$\mathbb{E}f(X_s(t)) - f^\star \leq \epsilon(s) + C(s)\mathrm{e}^{-\lambda_s t}, \tag{3}$$

where $f^\star$ denotes the global minimum of the objective function f, $\epsilon(s)$ denotes the risk at stationarity, and $C(s)$ depends on both the learning rate and the distribution of the initial x_0. Notably, we show that $\epsilon(s)$ decreases monotonically to zero as $s \to 0$, which is conducted from the temperature parameter in Raginsky et al. (2017). For any fixed time $T > 0$, this bound can be carried over to the discrete case by a uniform approximation between SGD and the LR-dependent SDE (2). Specifically, the term $C(s)\mathrm{e}^{-\lambda_s t}$ becomes $C(s)\mathrm{e}^{-\lambda_s k s}$, showing that the convergence is linear as well in the discrete regime. This is consistent with the numerical evidence from Figure 1 and Figure 2.

This convergence result sheds light on why SGD performs so well in many practical nonconvex problems. In particular, while GD can be trapped in a local minimum, SGD can efficiently escape it provided that the linear rate λ_s is not too small (this is the case if s is sufficiently large; see the second contribution). This superiority of SGD in the nonconvex setting must be attributed to the noise in the gradient and this implication is consistent with earlier work showing that stochasticity in gradients significantly accelerates the escape of saddle points for gradient-based methods Jin et al. (2017); Lee et al. (2016).

2. **Effect of learning rate on the nonconvex functions.** The first contribution stops short of saying anything about how λ_s depends on the learning rate s and the *geometry* of the objective f. Such an analysis is fundamental to an explanation of the differing effects of the learning rate in deep learning (nonconvex optimization) and convex optimization. In the current paper we show that if the objective f is a nonconvex function and satisfies certain regularity conditions, we have:[6]

$$\lambda_s \asymp \mathrm{e}^{-\frac{2H_f}{s}}, \tag{4}$$

for a certain value $H_f > 0$ that only depends on f. This expression for λ_s enables a concrete interpretation of the effect of learning rate in Figure 1. In brief, in the nonconvex setting, λ_s decreases to zero quickly as the learning rate s tends to zero. As a consequence, with a large learning rate s at the beginning, SGD converges rapidly to stationarity and the rate becomes smaller as the learning rate decreases.

5. Roughly speaking, stationarity refers to the distribution of $X_s(t)$ in the limit $t \to \infty$. See a more precise definition in Figure 3.
6. We write $a_m \asymp b_m$ if there exist positive constants c and c' such that $cb_m \leq a_m \leq c'b_m$ for all m.

For comparison, λ_s is equal to μ if f is μ-strongly convex for $\mu > 0$, regardless of the learning rate s. (In this case, the solution to the SDE converges to the global minimum with a learning rate of $1/t$ Hazan et al. (2008).) As such, the convergence behaviors of SGD are necessarily different between convex and nonconvex objectives. To appreciate this implication, we refer to Figure 3. Note that all four plots show

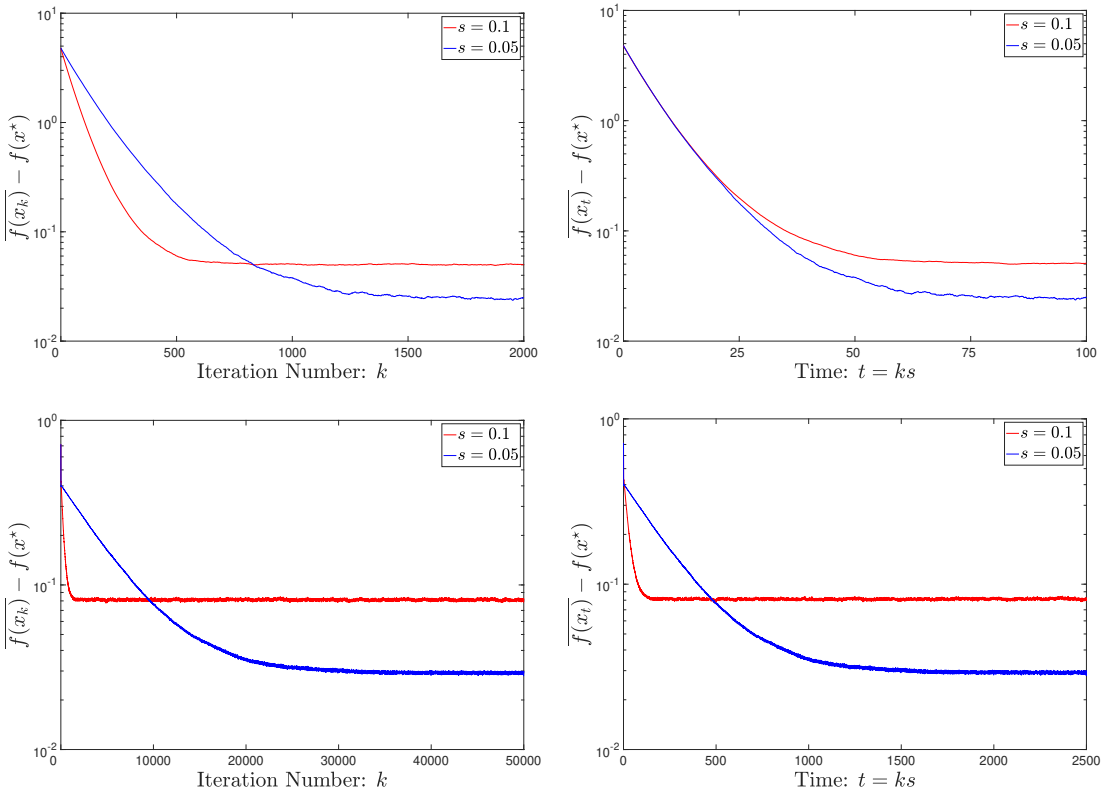

Figure 3: The dependence of the optimization dynamics of SGD on the learning rate *differs* between convex objectives and nonconvex objectives. The learning rate is set to either $s = 0.1$ or $s = 0.05$. The two top plots consider minimizing a convex function $f(x_1, x_2) = 5 \times 10^{-2} x_1^2 + 2.5 \times 10^{-2} x_2^2$, with an initial point $(8, 8)$, and the bottom plots consider minimizing a nonconvex function $f(x_1, x_2) = [(x_1 + 0.7)^2 + 0.1](x_1 - 0.7)^2 + (x_2 + 0.7)^2[(x_2 - 0.7)^2 + 0.1]$, with an initial point $(-0.9, 0.9)$. The gradient noise is drawn from the standard normal distribution. All results are averaged over 10000 independent replications.

that a larger learning rate gives rise to a larger stationary risk, as predicted by the monotonically increasing nature of ϵ with respect to s in (3). The most salient part of this figure is, however, shown in the right panel. Specifically, the right panel, which uses time t as the x-axis, shows that in the (strongly) convex setting the linear rate of the convergence is roughly the same between the two choices of learning rate, which is consistent with the result that λ_s is constant in the case of a strongly convex objective. In the nonconvex case (bottom right), however, the rate of convergence is more rapid

with the larger learning rate $s = 0.1$, which is implied by the fact that $\lambda_{0.1} > \lambda_{0.05}$. In stark contrast, the two plots in the left panel, which use the number k of iterations for the x-axis, are observed to have a larger rate of linear convergence with a larger learning rate. This is because in the k scale the rate $\lambda_s s$ of linear convergence always increases as s increases no matter if the objective is convex or nonconvex.

The mathematical tools that we bring to bear in analyzing the LR-dependent SDE (2) are as follows. We establish the linear convergence via a Poincaré-type inequality that is due to Villani Villani (2009). The asymptotic expression for the rate λ_s is proved by making use of the spectral theory of the Schrödinger operator or, more concretely, the Witten-Laplacian associated with the Fokker–Planck–Smoluchowski equation that governs the LR-dependent SDE. Different from the traditional probabilistic analysis, functional approaches are based on couplings, and the analysis based on the Schrödinger operator is based on the spectral theory of the operator, which is essentially an infinite-dimensional generalization of the finite-dimensional matrix. In other words, the analysis from operators reposes on an infinite-dimensional system, and generalizes the classical convergence analysis for a finite-dimensional dynamical system via the eigenvalues of the matrix Hirsch et al. (2012). Additionally, the spectral theory can be easily generalized to the momentum case. We believe that these tools will prove to be useful in theoretical analyses of other stochastic approximation methods.

1.2 Related work

Recent years have witnessed a surge of research devoted to explanations of the effectiveness of deep neural networks, with a particular focus on understanding how the learning rate affects the behavior of stochastic optimization. In Smith et al. (2017); Keskar et al. (2016), the authors uncovered various tradeoffs linking the learning rate and the mini-batch size. Moreover, Jastrzebski et al. (2017, 2018) related the learning rate to the generalization performance of neural networks in the early phase of training. This connection has been further strengthened by the demonstration that learning rate decay encourages SGD to learn features of increasing complexity Li et al. (2019b); You et al. (2019). From a topological perspective, Davis et al. (2019) establish connections between the learning rate and the sharpness of local minima. Empirically, deep learning models work well with non-decaying schedules such as cyclical learning rates Loshchilov and Hutter (2016); Smith (2017) (see also the review Sun (2019)), with recent theoretical justification Li and Arora (2019).

In a different direction, there has been a flurry of activity in using dynamical systems to analyze discrete optimization methods. For example, Su et al. (2016); Wibisono et al. (2016); Shi et al. (2018) derived ODEs for modeling Nesterov's accelerated gradient methods and used the ODEs to understand the acceleration phenomenon (see the review Jordan (2018)). In the stochastic setting, this approach has been recently pursued by various authors Chaudhari et al. (2018); Chaudhari and Soatto (2018); Mandt et al. (2016); Lee et al. (2016); Caluya and Halder (2019); Li et al. (2017) to establish various properties of stochastic optimization. As a notable advantage, the continuous-time perspective allows us to work without assumptions on the boundedness of the domain and gradients, as opposed to older analyses of SGD (see, for example, Hazan et al. (2008)).

Our work is motivated in part by the recent progress on Langevin dynamics, in particular in nonconvex settings Villani (2009); Pavliotis (2014); Helffer et al. (2004); Bovier et al. (2005). In relating to Langevin dynamics, s in the LR-dependent SDE can be thought of as the temperature parameter and, under certain conditions, this SDE has a stationary distribution given by the Gibbs measure, which is proportional to $\exp(-2f/s)$. Of particular relevance to the present paper from this perspective is a line of work that has considered the optimization properties of SGLD and analyzed its convergence rates Hwang (1980); Raginsky et al. (2017); Zhang et al. (2017). The LR-dependent SDE is formally similar to SGLD, in particular they both share the same Gibbs invariant distribution Raginsky et al. (2017). Linear convergence can be established for SGLD via the technique of the synchronous coupling Eberle (2016). Our approach provides an alternative to this line of work. The LR-dependent SDE is derived in our work as a surrogate for approximating SGD, and our analysis makes use of the Poincaré inequality under the Villani condition to obtain the L^2-distance of the probability densities instead of the 2-Wasserstein distance Raginsky et al. (2017). The advantages of our analysis hinge on the fact that it provides a concise and sharp delineation of the convergence rate based on the geometric properties of the objective function.

1.3 Organization

The remainder of the paper is structured as follows. In Section 2 we introduce basic assumptions and techniques employed throughout the paper. Section 3 develops our main theorems. In Section 4, we use the results of Section 3 to offer insights into the benefit of taking a larger initial learning rate followed by a sequence of decreasing learning rates in training neural networks. Section 5 formally proves the linear convergence (3) and Section 6 further specifies the rate of convergence (4). Technical details of the proofs are deferred to the appendices. We conclude the paper in Section 7 with a few directions for future research.

2. Preliminaries

Throughout this paper, we assume that the objective function f is infinitely differentiable in $\mathbb{R}^d$; that is, $f \in C^\infty(\mathbb{R}^d)$. We use $\|\cdot\|$ to denote the standard Euclidean norm.

Definition 1 (Confining condition Pavliotis (2014); Markowich and Villani (1999)) *A function f is said to be* confining *if it is infinitely differentiable and satisfies $\lim_{\|x\|\to+\infty} f(x) = +\infty$ and $\exp(-2f/s)$ is integrable for all $s > 0$:*

$$\int_{\mathbb{R}^d} \mathrm{e}^{-\frac{2f(x)}{s}} \mathrm{d}x < +\infty.$$

This condition is quite mild; it essentially requires that the function grows sufficiently rapidly when x is far from the origin. This condition is met, for example, when an ℓ_2 regularization term is added to the objective function f or, equivalently, weight decay is employed in the SGD update.

Next, we need to show that the LR-dependent SDE (2) with an arbitrary learning rate $s > 0$ admits a unique global solution under mild conditions on the objective f. We will

show in Section 3.3 that the solution to this SDE approximates the SGD iterates well. The formal description is shown rigorously in Proposition 8. Recall that the LR-dependent SDE (2) is

$$\mathrm{d}X_s = -\nabla f(X_s)\mathrm{d}t + \sqrt{s}\mathrm{d}W,$$

where the initial point $X_s(0)$ is distributed according to a probability density function ρ in $\mathbb{R}^d$, independent of the standard Brownian motion W. It is well known that the probability density $\rho_s(t,\cdot)$ of $X_s(t)$ evolves according to the LR-dependent Fokker–Planck–Smoluchowski equation

$$\frac{\partial \rho_s}{\partial t} = \nabla \cdot (\rho_s \nabla f) + \frac{s}{2}\Delta \rho_s, \tag{5}$$

with the boundary condition $\rho_s(0,\cdot) = \rho$. Here, $\Delta \equiv \nabla \cdot \nabla$ is the Laplacian. For completeness, in Appendix A.2 we derive this LR-dependent Fokker–Planck–Smoluchowski equation from the LR-dependent SDE (2) by Itô's formula. If the objective f satisfies the confining condition, then this equation admits a unique invariant Gibbs distribution that takes the form

$$\mu_s = \frac{1}{Z_s}\mathrm{e}^{-\frac{2f}{s}}. \tag{6}$$

The proof of uniqueness is shown in Appendix A.3. The normalization factor is $Z_s = \int_{\mathbb{R}^d} \mathrm{e}^{-\frac{2f}{s}}\mathrm{d}x$. Taking any initial probability density $\rho_s(0,\cdot) \equiv \rho$ in $L^2(\mu_s^{-1})$ (a measurable function g is said to belong to $L^2(\mu_s^{-1})$ if $\|g\|_{\mu_s^{-1}} := \left(\int_{\mathbb{R}^d} g^2 \mu_s^{-1}\mathrm{d}x\right)^{\frac{1}{2}} < +\infty$), we have the following guarantee:

Lemma 2 (Existence and uniqueness of the weak solution) *For any confining function f and any initial probability density $\rho \in L^2(\mu_s^{-1})$, the LR-dependent SDE* (2) *admits a weak solution whose probability density in $C^1\left([0,+\infty), L^2(\mu_s^{-1})\right)$ is the unique solution to the LR-dependent Fokker–Planck–Smoluchowski equation* (5)*.*

The proof of Lemma 2 can be obtained by Harnack's inequality, a classical approach using a second-order elliptic operator, as described in Bogachev et al. (2009). We present an alternative proof of Lemma 2 based on the spectral theory of the Schrödinger operator in Appendix A.4. We also present a companion result in Lemma 10 in Section 5, which shows that the probability density $\rho_s(t,\cdot)$ converges to the Gibbs distribution as $t \to \infty$. Finally, we need a condition that is due to Villani for the development of our main results in the next section.

Definition 3 (Villani condition Villani (2009)) *A confining function f is said to satisfy the* Villani condition *if $\|\nabla f(x)\|^2/s - \Delta f(x) \to +\infty$ as $\|x\| \to +\infty$ for all $s > 0$.*

This condition amounts to saying that the gradient has a sufficiently large squared norm compared with the Laplacian of the function. Strictly speaking, some loss functions used for training neural networks might not satisfy this condition. However, the Villani condition does not look as stringent as it appears since this condition is essentially concerned with the function at infinity. In this paper, we use the Villani condition to derive the Poincaré inequality and the discrete spectrum of the Witten-Laplacian. There are alternatives to the

Villani condition; see (Bakry et al., 2008, Corollary 1.6). For example, we can replace the Villani condition with the following condition

$$\frac{\langle x, \nabla f(x)\rangle}{|x|} \to +\infty$$

as $\|x\| \to +\infty$. However, it is unknown whether these conditions lead to the result that the spectrum of Witten-Laplacian is discrete.

3. Main Results

In this section, we state our main results. In brief, in Section 3.1 we show linear convergence to stationarity for SGD in its continuous formulation, the LR-dependent SDE. In Section 3.2, we derive a quantitative expression of the rate of linear convergence and study the difference in the behavior of SGD in the convex and nonconvex settings. This distinction is further elaborated in Section 3.3 by carrying over the continuous-time convergence guarantees to the discrete case. Finally, Section 3.4 offers an exposition of the theoretical results in the univariate case. Proofs of the results presented in this section are deferred to Section 5 and Section 6.

3.1 Linear convergence

In this subsection we are concerned with the expected excess risk, $\mathbb{E}f(X_s(t)) - f^\star$. Recall that $f^\star = \inf_x f(x)$.

Theorem 1 *Let f satisfy both the confining condition and the Villani condition. Then there exists $\lambda_s > 0$ for any learning rate $s > 0$ such that the expected excess risk satisfies*

$$\mathbb{E}f(X_s(t)) - f^\star \le \epsilon(s) + D(s)\mathrm{e}^{-\lambda_s t}, \tag{7}$$

for all $t \ge 0$. Here $\epsilon(s) = \epsilon(s; f) \ge 0$ is a strictly increasing function of s depending only on the objective function f, and $D(s) = D(s; f, \rho) \ge 0$ depends only on s, f, and the initial distribution ρ.

Briefly, the proof of this theorem is based on the following decomposition of the excess risk:

$$\mathbb{E}f(X_s(t)) - f^\star = \mathbb{E}f(X_s(t)) - \mathbb{E}f(X_s(\infty)) + \mathbb{E}f(X_s(\infty)) - f^\star,$$

where we informally use $\mathbb{E}f(X_s(\infty))$ to denote $\mathbb{E}_{X\sim\mu_s} f(X)$ in light of the fact that $X_s(t)$ converges weakly to μ_s as $t \to +\infty$ (see Lemma 10). The question is thus to quantify how fast $\mathbb{E}f(X_s(t)) - \mathbb{E}f(X_s(\infty))$ vanishes to zero as $t \to \infty$ and how the excess risk at stationarity $\mathbb{E}f(X_s(\infty)) - f^\star$ depends on the learning rate. The following two propositions address these two questions. Recall that $\rho \in L^2(\mu_s^{-1})$ is the probability density of the initial iterate in SGD.

Proposition 4 *Under the assumptions of Theorem 1, there exists $\lambda_s > 0$ for any learning rate s such that*

$$|\mathbb{E}f(X_s(t)) - \mathbb{E}f(X_s(\infty))| \le C(s)\,\|\rho - \mu_s\|_{\mu_s^{-1}}\,\mathrm{e}^{-\lambda_s t},$$

for all $t \geq 0$, where the constant $C(s) > 0$ depends only on s and f, and where

$$\|\rho - \mu_s\|_{\mu_s^{-1}} = \left(\int_{\mathbb{R}^d} (\rho - \mu_s)^2 \, \mu_s^{-1} \mathrm{d}x \right)^{\frac{1}{2}}$$

measures the gap between the initialization and the stationary distribution.

Loosely speaking, it takes $O(1/\lambda_s)$ time to converge to stationarity. In relating to Theorem 1, $D(s)$ can be set to $C(s) \|\rho - \mu_s\|_{\mu_s^{-1}}$. Notably, the proof of Proposition 4 shall reveal that $C(s)$ increases as s increases. Turning to the analysis of the second term, $\mathbb{E}f(X_s(\infty)) - f^\star$, we henceforth write $\epsilon(s) := \mathbb{E}f(X_s(\infty)) - f^\star$.

Proposition 5 *Under the assumptions of Theorem 1, the excess risk at stationarity, $\epsilon(s)$, is a strictly increasing function of s. Moreover, for any $S > 0$, there exists a constant A that depends only on S and f and satisfies*

$$\epsilon(s) \equiv \mathbb{E}f(X_s(\infty)) - f^\star \leq As,$$

for any learning rate $0 < s \leq S$.

The two propositions are proved in Section 5. The proof of Theorem 1 is a direct consequence of Proposition 4 and Proposition 5. More precisely, the two propositions taken together give

$$\mathbb{E}f(X_s(t)) - f^\star \leq O(s) + C(s)\mathrm{e}^{-\lambda_s t}, \tag{8}$$

for a bounded learning rate s. Note that a dimension-dependent upper bound of $O(ds)$ is provided for $\epsilon(s)$ in (Raginsky et al., 2017, Section 3.5). This estimate is obtained by evaluating both the second moment of the invariant distribution via the Euclidean 2-Wasserstein distance and the integral constant based on the global gradient Lipschitz condition. However, it is worth noting that the global gradient Lipschitz condition may not be practical in real-world scenarios. In line with Theorem 2, the constant A in this case is also dependent on the geometry of f; for more details see Section 5.2.

Taken together, these results offer insights into the phenomena observed in Figure 1. In particular, Proposition 4 states that, from the continuous-time perspective, the risk of SGD with a constant learning rate applied to a (nonconvex) objective function converges to stationarity at a *linear* rate. Moreover, Proposition 5 demonstrates that the excess risk at stationarity decreases as the learning rate s tends to zero. This is in agreement with the numerical experiments illustrated in Figure 1, Figure 2, and Figure 3. For comparison, this property is not observed in GD and SGLD. The following result gives the time complexity of SGD in its continuous-time formulation.

Corollary 6 *Under the assumptions of Proposition 5, for any $\epsilon > 0$, if the learning rate $s \leq \min\{\epsilon/(2A), S\}$ and $t \geq \frac{1}{\lambda_s} \log \frac{2C(s)\|\rho - \mu_s\|_{\mu_s^{-1}}}{\epsilon}$, then*

$$\mathbb{E}f(X_s(t)) - f^\star \leq \epsilon.$$

3.2 The rate of linear convergence

We now turn to the key issue of understanding how the linear rate λ_s depends on the learning rate. In this subsection, we show that for certain objective functions, λ_s admits a simple expression that allows us to interpret how the convergence rate depends on the learning rate.

We begin by considering a strongly convex function. Recall the definition of strong convexity: for $\mu > 0$, a function f is μ-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2,$$

for all x, y. Equivalently, f is μ-strong convex if all eigenvalues of its Hessian $\nabla^2 f(x)$ are greater than or equal to μ for all x (note that here f is assumed to be infinitely differentiable). As is clear, a strongly convex function satisfies the confining condition. In Appendix B.1, we prove the following proposition by making use of a Poincaré-type inequality, the Bakry–Emery theorem Bakry et al. (2013).

Proposition 7 *In addition to the assumptions of Theorem 1, assume that the objective f is a μ-strongly convex function. Then, λ_s in (7) satisfies $\lambda_s = \mu$.*

We turn to the more challenging setting where f is *nonconvex*. Let us refer to the objective f as a *Morse function* if its Hessian has full rank at any critical point x (that is, $\nabla f(x) = 0$).[7]

Theorem 2 *In addition to the assumptions of Theorem 1, assume that the objective f is a Morse function and has at least two local minima.*[8] *Then the constant λ_s in (7) satisfies*

$$\lambda_s = (\alpha + o(s))\mathrm{e}^{-\frac{2H_f}{s}}, \tag{9}$$

for $0 < s \leq s_0$, where $s_0 > 0, \alpha > 0$, and $H_f > 0$ are constants that all depend only on f.

The proof of this result relies on tools in the spectral theory of Schrödinger operators and is deferred to Section 6. From now on, we call λ_s in (7) the *exponential decay constant*. To obviate any confusion, $o(s)$ in Theorem 2 stands for a quantity that tends to zero as $s \to 0$, and the precise expression for H_f shall be given in Section 6, with a simple example provided in Section 3.4. To leverage Theorem 2 for understanding the phenomena discussed in Section 1, however, it suffices to recognize the fact that H_f is completely determined by f. Moreover, we remark that while Theorem 1 shows that λ_s exists for any learning rate, the present theorem assumes a bounded learning rate.

The key implication of this result is that the rate of convergence is highly contingent upon the learning rate s: the exponential decay constant increases as the learning rate s increases. Accordingly, the linear convergence to stationarity established in Section 3.1 is faster if s is larger, and, by recognizing the exponential dependence of λ_s on s, the convergence would

7. See Section 6.2 for a discussion of Morse functions. Note that (infinitely differentiable) strongly convex functions are Morse functions.
8. We call x a local minimum of f if $\nabla f(x) = 0$ and the Hessian $\nabla^2 f(x)$ is positive definite. By convention, in this paper a global minimum is also considered a local minimum.

be very slow if the learning rate s is very small. For example, if $H_f = 0.05$, setting $s = 0.1$ and $s = 0.001$ gives

$$\frac{\lambda_{0.1}}{\lambda_{0.001}} \approx \frac{\mathrm{e}^{-1}}{\mathrm{e}^{-100}} = 9.889 \times 10^{42}.$$

Moreover, as we will see clearly in Section 6, λ_s is completely determined by the *geometry* of f. In particular, it does not depend on the probability distribution of the initial point or the dimension d given that the constant H_f has no direct dependence on the dimension d. For comparison, the linear rate in the nonconvex case is shown by Theorem 2 to depend on the learning rate s, while the linear rate of convergence stays constant regardless of s if the objective is strongly convex. This fundamental distinction between the convex and nonconvex settings enables an interpretation of the observation brought up in Figure 1, in particular the right panel of Figure 3. More precisely, with time t being the x-axis, SGD with a larger learning rate leads to a faster convergence rate in the nonconvex setting, while for the (strongly) convex setting the convergence rate is independent of the learning rate. For further in-depth discussion of the implications of Theorem 2, see Section 4.

3.3 Discretization

In this subsection, we carry over the results developed from the continuous perspective to the discrete regime. In addition to assuming that the objective function f satisfies the Villani condition, satisfies the confining condition, and is a Morse function, we also now assume f to be L-smooth; that is, f has L-Lipschitz continuous gradients in the sense that $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$ for all x, y. Moreover, we restrict the learning rate s to be no larger than $1/L$. The following proposition is the key theoretical tool that allows translation to the discrete regime.

Proposition 8 *For any L-smooth objective f and any initialization $X_s(0)$ drawn from a probability density $\rho \in L^2(\mu_s^{-1})$, the LR-dependent SDE (2) has a unique global solution X_s in expectation; that is, $\mathbb{E}X_s(t)$ as a function of t in $C^1([0, +\infty); \mathbb{R}^d)$ is unique. Moreover, there exists $B(T) > 0$ such that the SGD iterates x_k satisfy*

$$\max_{0 \le k \le T/s} |\mathbb{E}f(x_k) - \mathbb{E}f(X_s(ks))| \le B(T)s,$$

for any fixed $T > 0$.

We note that there exists a sharp bound on $B(T)$ in Bally and Talay (1996). For completeness, we also remark that the convergence can be strengthened to the strong sense:

$$\max_{0 \le k \le T/s} \mathbb{E}\,\|x_k - X_s(ks)\| \le B'(T)s.$$

This result has appeared in Mil'shtein (1975); Talay (1982); Pardoux and Talay (1985); Talay (1984); Kloeden and Platen (1992) and we provide a self-contained proof in Appendix B.2.

We now state the main result of this subsection.

Theorem 3 *In addition to the assumptions of Theorem 1, assume that f is L-smooth. Then, for any $T > 0$, the iterates of SGD with learning rate $0 < s \leq 1/L$ satisfy*

$$\mathbb{E}f(x_k) - f^\star \leq (A + B(T))s + C \left\|\rho - \mu_s\right\|_{\mu_s^{-1}} \mathrm{e}^{-s\lambda_s k}, \tag{10}$$

for all $k \leq T/s$, where λ_s is the exponential decay constant in (7), A as in Proposition 5 depends only on $1/L$ and f, $C = C_{1/L}$ is as in Proposition 4, and $B(T)$ depends only on the time horizon T and the Lipschitz constant L.

Theorem 3 follows as a direct consequence of Theorem 1 and Proposition 8. Note that if f is a Morse function with at least two local minima, then λ_s appearing in (10) is given by (9), and if f is μ-strongly convex then $\lambda_s = \mu$. As earlier in the continuous-time formulation, we also mention that the dimension parameter d is not an essential parameter for characterizing the rate of linear convergence. In relating to Figure 3, note that its left panel with k being the x-axis shows a faster linear convergence of SGD when using a larger learning rate, regardless of convexity or nonconvexity of the objective. This is because the linear rate $s\lambda_s$ in (10) is always an increasing function of s even for the strongly convex case, where λ_s itself is constant.

3.4 A one-dimensional example

In this section we provide some intuition for the theoretical results presented in the preceding subsections. Our priority is to provide intuition rather than rigor. Consider the simple example of f presented in Figure 4, which has a global minimum $x^\star$, a local minimum $x^\bullet$, and a local maximum x°.[9] We use this toy example to gain insight into the expression (9) for the exponential decay constant λ_s; deferring the rigorous derivation of this number in the general case to Section 6.

From (7) it appears that the LR-dependent SDE (2) takes about $O(1/\lambda_s)$ time to achieve approximate stationarity. Intuitively, for the specific function in Figure 4, the bottleneck in achieving stationarity is to pass through the local maximum x°. Now, we show that it takes about $O(1/\lambda_s)$ time to pass x° from the local minimum $x^\bullet$. For simplicity, write

$$f(x) = \frac{\theta}{2}(x - x^\bullet)^2 + g(x),$$

where $g(x) = f(x^\bullet)$ stays constant if $x \leq x^\circ - \nu$ for a very small positive ν and $\theta > 0$. Accordingly, the LR-dependent SDE (2) is reduced to the Ornstein–Uhlenbeck process,

$$\mathrm{d}X_s = -\theta(X_s - x^\bullet)\mathrm{d}t + \sqrt{s}\mathrm{d}W,$$

before hitting x°. Denote by τ_{x° the first time the Ornstein–Uhlenbeck process hits x°. It is well known that the hitting time obeys

$$\mathbb{E}\tau_{x^\circ} \approx \frac{\sqrt{\pi s}}{(x^\circ - x^\bullet)\theta\sqrt{\theta}} \mathrm{e}^{\frac{2}{s}\cdot\frac{1}{2}\theta(x^\circ - x^\bullet)^2} \approx \frac{\sqrt{\pi s}}{(x^\circ - x^\bullet)\theta\sqrt{\theta}} \mathrm{e}^{\frac{2H_f}{s}}, \tag{11}$$

9. We can also regard x° as a saddle point in the sense that the Hessian at this point has one negative eigenvalue. See Section 6.2 for more discussion.

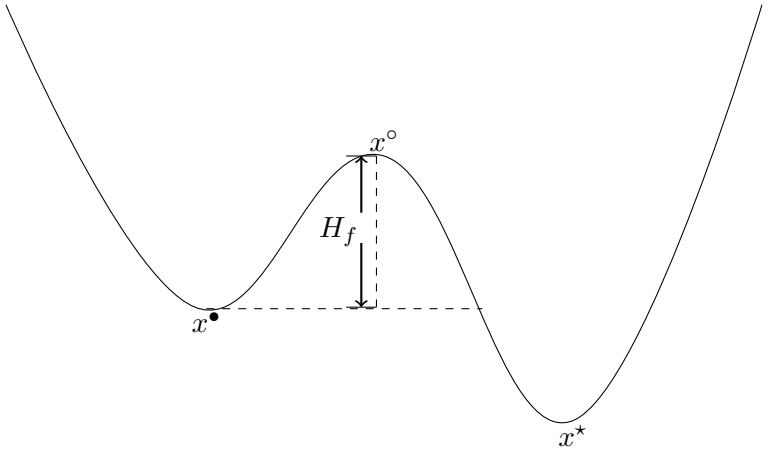

Figure 4: A one-dimensional nonconvex function f. The height difference between x° and $x^\bullet$ in this special case is the Morse saddle barrier H_f. See the formal definition in Definition 20.

where $H_f := f(x^\circ) - f(x^\bullet) \approx f(x^\circ) - g(x^\circ) = \frac{1}{2}\theta(x^\circ - x^\bullet)^2$. This number, which we refer to as the *Morse saddle barrier*, is the difference between the function values at the local maximum x° and the local minimum $x^\bullet$ in our case. As an implication of (11), the continuous-time formulation of SGD takes time (at least) of the order $\mathrm{e}^{(1+o(1))\frac{2H_f}{s}}$ to achieve approximate stationarity. This is consistent with the exponential decay constant λ_s given in (9).

In passing, we remark that the discussion above can be made rigorous by invoking the theory of the Kramers' escape rate, which shows that for this univariate case the hitting time satisfies

$$\mathbb{E}\tau_{x^\circ} = (1 + o(1))\frac{\pi}{\sqrt{-f''(x^\bullet)f''(x^\circ)}}\mathrm{e}^{\frac{2H_f}{s}}.$$

See, for example, Freidlin and Wentzell (2012); Pavliotis (2014). Furthermore, we demonstrate the view from the theory of *viscosity solutions* and *singular perturbations* in Appendix B.3.

4. Why Learning Rate Decay?

As a widely used technique for training neural networks, learning rate decay refers to taking a large learning rate initially and then progressively reducing it during the training process. This technique has been observed to be highly effective especially in the minimization of nonconvex objective functions using stochastic optimization methods, with a very recent strand of theoretical effort aiming at understanding its benefits You et al. (2019); Li et al. (2019b). In this section, we offer a new and crisp explanation by leveraging the results in Section 3. To highlight the intuition, we primarily work with the continuous-time formulation of SGD.

For purposes of illustration, Figure 5 presents numerical examples for this technique where the learning rate is set to 0.1 or 0.05. This figure clearly demonstrates that SGD

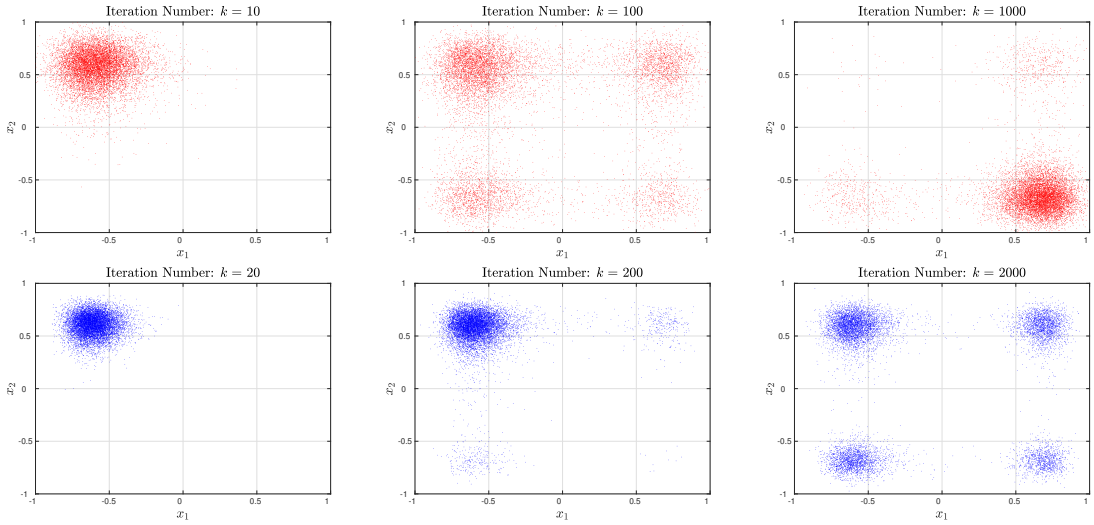

Figure 5: Scatter plots of the iterates $x_k \in \mathbb{R}^2$ of SGD for minimizing the nonconvex function in Figure 3. This function has four local minima, of which the bottom right one is the global minimum. Each column corresponds to the same value of $t = ks$, and the first row and second row correspond to learning rates 0.1 and 0.05, respectively. The gradient noise is drawn from the standard normal distribution. Each plot is based on 10000 independent SGD runs using the noise generator "state 1-10000" in Matlab2019b, starting from an initial point $(-0.9, 0.9)$.

with a larger learning rate converges much faster to the global minimum than SGD with a smaller learning rate. This comparison reveals that a large learning rate would render SGD able to quickly explore the landscape of the objective function and efficiently escape bad local minima. On the other hand, a larger learning rate would prevent SGD iterates from concentrating around a global minimum, leading to substantial suboptimality. This is clearly illustrated in Figure 6. As suggested by the heuristic work on learning rate decay, we see that it is important to decrease the learning rate to achieve better optimization performance whenever the iterates arrive near a local minimum of the objective function.

Despite its intuitive plausibility, the exposition above stops short of explaining why nonconvexity of the objective is crucial to the effectiveness of learning rate decay. Our results in Section 3, however, enable a concrete and crisp understanding of the vital importance of nonconvexity in this setting. Motivated by (8), we consider an idealized risk function of the form $R(t) = as + (b - as)\mathrm{e}^{-\lambda_s t}$, with λ_s set to $\mathrm{e}^{-c/s}$, where a, b, and c are positive constants for simplicity as opposed to the non-constants in the upper bound in (7). This function is plotted in Figure 7, with two quite different learning rates, $s_1 = 0.1$ and $s_2 = 0.001$, as an implementation of learning rate decay. When the learning rate is $s_1 = 0.1$, from the right panel of Figure 7, we see that rough stationarity is achieved at time $t = ks \approx 25$; thus, the number of iterations $k_{0.1} \approx 25/s = 250$. In the case of $s = 0.001$, from the left panel of Figure 7, we see now it requires $ks \approx 2.5 \times 10^{44}$ to reach rough stationarity, leading to

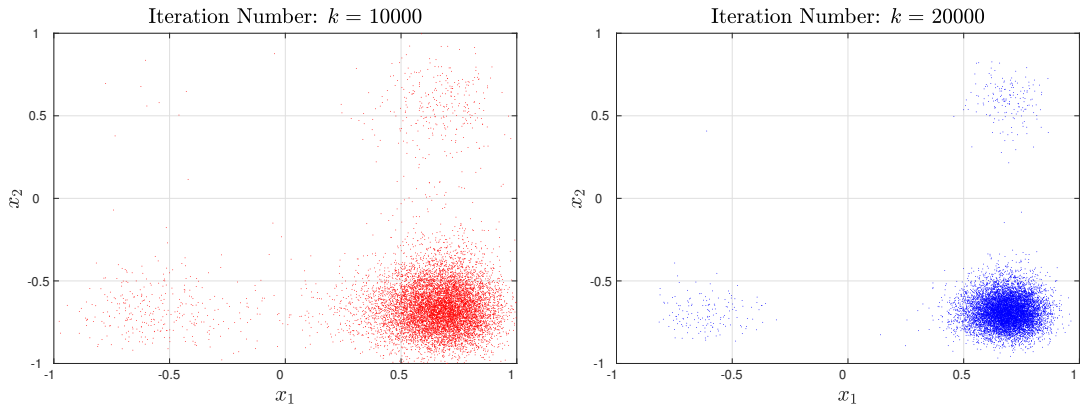

Figure 6: The same setting as in Figure 5. Both plots correspond to the same value of $t = ks = 1000$.

$k_{0.001} \approx 2.5 \times 10^{47}$. This gives

$$\frac{k_{0.001}}{k_{0.1}} \approx 10^{45}.$$

In contrast, the sharp dependence of k_s on the learning rate s is not seen for strongly convex functions, because $\lambda_s = \mu$ stays constant as the learning rate s varies. Following the preceding example, we have

$$\frac{k_{0.001}}{k_{0.1}} \approx 10^2.$$

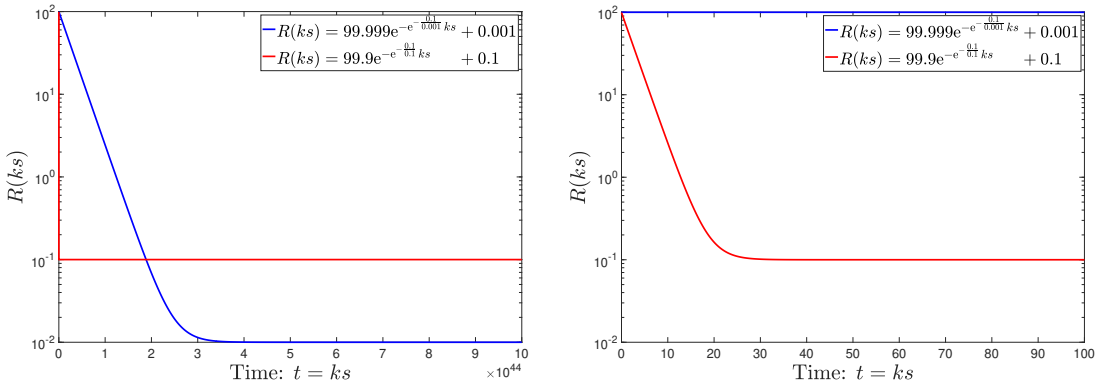

Figure 7: Idealized risk function of the form $R(t) = as + (b - as)e^{-e^{\frac{c}{s}t}}$ with the identification $t = ks$, which is adapted from (8). The parameters are set as follows: $a = 1, b = 100, c = 0.1$, and the learning rate is $s = 0.1$ or 0.001. The right plot is a locally enlarged image of the left.

While a large initial learning rate helps speed up the convergence, Figure 7 also demonstrates that a larger learning rate leads to a larger value of the excess risk at stationarity,

$\epsilon(s) \equiv \mathbb{E}f(X_s(\infty)) - f^\star$, which is indeed the claim of Proposition 5. Leveraging Proposition 4, we show below why annealing the learning rate at some point would improve the optimization performance. To this end, for any fixed learning rate s, consider a stopping time T_s^δ that is defined as

$$T_s^\delta := \inf_t \left\{ |\mathbb{E}f(X_s(t)) - \mathbb{E}f(X_s(\infty))| \le \delta\epsilon(s) \right\},$$

for a small $\delta > 0$. In words, the LR-dependent SDE (2) at time T_s^δ is approximately stationary since its risk $\mathbb{E}f(X_s(t)) - f^\star$ is mainly comprised of the excess risk at stationarity $\epsilon(s)$, with a total risk of no more than $(1+\delta)\epsilon(s)$. From Proposition 4 it follows that (recall that ρ is the initial distribution):

$$T_s^\delta \le \frac{1}{\lambda_s} \log \frac{C(s) \left\| \rho - \mu_s \right\|_{\mu_s^{-1}}}{\delta\epsilon(s)} = \frac{\mathrm{e}^{\frac{2H_f}{s}}}{\gamma + o(s)} \log \frac{C(s) \left\| \rho - \mu_s \right\|_{\mu_s^{-1}}}{\delta\epsilon(s)}. \tag{12}$$

In addition to taking a large s, an alternative way to make T_s^δ small is to have an initial distribution ρ that is close to the stationary distribution μ_s. This can be achieved by using the technique of learning rate decay. More precisely, taking a larger learning rate s_1 for a while, at the end the distribution of the iterates is approximately the stationary distribution μ_{s_1}, which serves as the initial distribution for SGD with a smaller learning rate s_2 in the second phase. Taking $\rho \approx \mu_{s_1}$, the factor $\left\| \rho - \mu_s \right\|_{\mu_s^{-1}}$ in (12) for the second phase of learning rate decay is approximately

$$\left\| \mu_{s_1} - \mu_{s_2} \right\|_{\mu_{s_2}^{-1}} = \left(\int (\mu_{s_1} - \mu_{s_2})^2 \mu_{s_2}^{-1} \mathrm{d}x \right)^{\frac{1}{2}} = \left(\int \frac{\mu_{s_1}^2}{\mu_{s_2}} \mathrm{d}x - 1 \right)^{\frac{1}{2}}. \tag{13}$$

Both μ_{s_1} and μ_{s_2} are decreasing functions of f and, therefore, have the same modes. As $s_1 \to 0$, both μ_{s_1} and μ_{s_2} tend to $\delta(x - x^\star)$, thereby implying $\mu_{s_1}/\mu_{s_2} \to 1$. As a consequence, the integral of $\mu_{s_1}^2/\mu_{s_2}$ minus one is small by appealing to the rearrangement inequality, thereby leading to fast convergence of SGD with learning rate s_2 to the stationary risk $\epsilon(s_2)$. In contrast, $\left\| \rho - \mu_{s_2} \right\|_{\mu_{s_2}^{-1}}$ would be much larger for a general random initialization ρ. Put simply, SGD with learning rate s_2 cannot achieve a risk of approximately $\epsilon(s_2)$ given the same number of iterations *without* the warm-up stage using learning rate s_1. See Figure 8 for an illustration.

$$\rho \xrightarrow{\text{large learning rate } s_1} \approx \mu_{s_1} \xrightarrow{\text{small learning rate } s_2} \approx \mu_{s_2}$$

Figure 8: Learning rate decay. The first phase uses a larger learning rate s_1, at the end of which the SGD iterates are approximately distributed as μ_{s_1}. The second phase uses a smaller learning rate s_2 and at the end the distribution of the SGD iterates roughly follows μ_{s_2}.

In Chiang et al. (1987), the concept of simulated annealing is introduced to the diffusion process. It is equivalent to the time-decay learning rate as $s = c/\log t$ in the LR-dependent SDE. Through probabilistic analysis, Chiang et al. (1987) derives that the linear rate is $t^{-c'}$

and the invariant distribution is $\exp(-2f(x)\log t/c)$.[10] Although the invariant distribution concentrates on the global minima as t approaches infinity, the linear rate decays rapidly, resulting in extremely slow convergence of the distribution. However, it is worth noting that the process described in Chiang et al. (1987) is a single-phase process. In practice, the phenomenon generated by the learning rate decaying piecewise is likely a two-phase process. The first phase involves global convergence with the learning rate $\exp(-2H_f/s_1)$, while the second phase is likely the local convergence with the learning rate μ. This is because most of the invariant distribution $\exp(-2f(x)/s_1)$ concentrates on the neighborhood of the global minima. We note also that Kushner (1987) provides a derivation of both the mean escape time and the mean transition time using the theory of large deviations for the discrete case.

5. Proof of the Linear Convergence

In this section, we prove Proposition 4 and Proposition 5, leading to a complete proof of Theorem 1.

5.1 Proof of Proposition 4

To better appreciate the linear convergence of the LR-dependent SDE (2), as established in Proposition 4, we start by showing the convergence to stationarity without a rate. In fact, this intermediate result constitutes a necessary step in the proof of Proposition 4. The techniques presented in this section are standard in the literature (see, for example, Villani (2009); Pavliotis (2014)).

Convergence without a rate. Recall that we use ρ to denote the initial probability density in $L^2(\mu_s^{-1})$. Superficially, it seems that the most natural space for probability densities is $L^1(\mathbb{R}^d)$. However, it is mathematically convenient to work on an inner product space as opposed to a general Banach space to prove convergence results for the LR-dependent SDE. Indeed, studying densities in $L^2(\mu_s^{-1})$ is a common strategy. Formally, the following result says that any (nonnegative) function in $L^2(\mu_s^{-1})$ can be normalized to be a density function. The proof of this simple lemma is shown in Appendix C.1.

Lemma 9 *Let f satisfy the confining condition. Then, $L^2(\mu_s^{-1})$ is a subset of $L^1(\mathbb{R}^d)$.*

The following result shows that the solution to the LR-dependent SDE converges to stationarity in terms of the dynamics of its probability densities over time.

Lemma 10 *Let f satisfy the confining condition and denote the initial distribution as $\rho \in L^2(\mu_s^{-1})$. Then, the unique solution $\rho_s(t,\cdot) \in C^1\left([0,+\infty), L^2(\mu_s^{-1})\right)$ to the Fokker–Planck–Smoluchowski equation (5) converges in $L^2(\mu_s^{-1})$ to the Gibbs invariant distribution μ_s, which is specified by (6).*

10. The constant c' mentioned here is used to distinguish it from the previous constant c. It serves as a separate identifier to avoid confusion or ambiguity.

Proof [Proof of Lemma 10] We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\rho_s(t,\cdot)-\mu_s\right\|_{\mu_s^{-1}}^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\left(h_s(t,x)-1\right)^2\mathrm{d}\mu_s \\ &= 2\int_{\mathbb{R}^d}(h_s-1)\mathscr{L}_s(h_s-1)\mathrm{d}\mu_s,\end{aligned}$$

where the last equality is due to (15). Next, we proceed by making use of Lemma 11:

$$\begin{aligned}2\int_{\mathbb{R}^d}(h_s-1)\mathscr{L}_s(h_s-1)\mathrm{d}\mu_s &= -s\int_{\mathbb{R}^d}\nabla(h_s-1)\cdot\nabla(h_s-1)\mathrm{d}\mu_s \\ &= -s\int_{\mathbb{R}^d}\|\nabla h_s\|^2\mathrm{d}\mu_s \le 0. \end{aligned} \tag{14}$$

Thus, $\left\|\rho_s(t,\cdot)-\mu_s\right\|_{\mu_s^{-1}}^2$ is a strictly decreasing function, decreasing asymptotically towards the equilibrium state

$$\int_{\mathbb{R}^d}\|\nabla h_s\|^2\mathrm{d}\mu_s = 0.$$

This equality holds, however, only if $h_s(t,\cdot)$ is constant. Because both $\rho_s(t,\cdot)$ and μ_s are probability densities, this case must imply that $h_s(t,\cdot)\equiv 1$; that is, $\rho_s(t,\cdot)\equiv\mu_s$. Therefore, $\rho_s(t,\cdot)\in C^1\left([0,+\infty),L^2(\mu_s^{-1})\right)$ converges to the Gibbs invariant distribution μ_s in $L^2(\mu_s^{-1})$. ■

Note that the existence and uniqueness of $\rho_s(t,\cdot)$ is ensured by Lemma 2. The convergence guarantee on $\rho_s(t,\cdot)$ in Lemma 10 relies heavily on the following lemma (Lemma 11). This preparatory lemma introduces the transformation

$$h_s(t,\cdot)=\rho_s(t,\cdot)\mu_s^{-1}\in C^1\left([0,+\infty),L^2(\mu_s)\right),$$

which allows us to work in the space $L^2(\mu_s)$ in place of $L^2(\mu_s^{-1})$ (a measurable function g is said to belong to $L^2(\mu_s)$ if $\|g\|_{\mu_s}:=\left(\int_{\mathbb{R}^d}g^2\mathrm{d}\mu_s\right)^{\frac{1}{2}}<+\infty$).[11] It is not hard to show that h_s satisfies the following equation

$$\frac{\partial h_s}{\partial t}=-\nabla f\cdot\nabla h_s+\frac{s}{2}\Delta h_s, \tag{15}$$

with the initial distribution $h_s(0,\cdot)=\rho\mu_s^{-1}\in L^2(\mu_s)$. The linear operator

$$\mathscr{L}_s=-\nabla f\cdot\nabla+\frac{s}{2}\Delta \tag{16}$$

has a crucial property, as stated in the following lemma, whose proof is provided in Appendix C.2.

Lemma 11 *The linear operator $\mathscr{L}_s$ in* (16) *is self-adjoint and nonpositive in $L^2(\mu_s)$. Explicitly, for any g_1,g_2, this operator obeys*

$$\int_{\mathbb{R}^d}(\mathscr{L}_s g_1)g_2\mathrm{d}\mu_s=\int_{\mathbb{R}^d}g_1\mathscr{L}_s g_2\mathrm{d}\mu_s=-\frac{s}{2}\int_{\mathbb{R}^d}\nabla g_1\cdot\nabla g_2\mathrm{d}\mu_s.$$

11. Here, $\mathrm{d}\mu_s$ stands for the probability measure $\mathrm{d}\mu_s\equiv\mu_s\mathrm{d}x=\frac{1}{Z_s}\exp(-2f/s)\mathrm{d}x$.

Linear convergence. We turn to the proof of linear convergence. We first state a lemma which serves as a fundamental tool for us to prove a linear rate of convergence for Proposition 4.

Lemma 12 (Theorem A.1 in Villani (2009)) *If f satisfies both the confining condition and the Villani condition, then there exists $\lambda_s > 0$ such that the measure $\mathrm{d}\mu_s$ satisfies the following Poincaré-type inequality*

$$\int_{\mathbb{R}^d} h^2 \mathrm{d}\mu_s - \left(\int_{\mathbb{R}^d} h \mathrm{d}\mu_s \right)^2 \leq \frac{s}{2\lambda_s} \int_{\mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s,$$

for any h such that the integrals are well defined.

For completeness, we provide a proof of this Poincaré-type inequality in Appendix C.3. For comparison, the usual Poincaré inequality is put into use for a bounded domain, as opposed to the entire Euclidean space as in Lemma 12. In addition, while the constant in the Poincaré inequality in general depends on the dimension (see, for example, (Evans, 2010, Theorem 1, Chapter 5.8)), λ_s in Lemma 12 is completely determined by geometric properties of the objective f. See details in Section 6.

Importantly, Lemma 12 allows us to obtain the following lemma, from which the proof of Proposition 4 follows readily. The proof of this lemma is given at the end of this subsection.

Lemma 13 *Under the assumptions of Proposition 4, $\rho_s(t, \cdot)$ converges to the Gibbs invariant distribution μ_s in $L^2(\mu_s^{-1})$ at the rate*

$$\left\| \rho_s(t, \cdot) - \mu_s \right\|_{\mu_s^{-1}} \leq \mathrm{e}^{-\lambda_s t} \left\| \rho - \mu_s \right\|_{\mu_s^{-1}}. \tag{17}$$

Proof [Proof of Proposition 4] Using Lemma 13, we get

$$\begin{aligned}
\left| \mathbb{E} f(X_s(t)) - \mathbb{E} f(X(\infty)) \right| &= \left| \int_{\mathbb{R}^d} (f(x) - f^\star) \left(\rho_s(t, x) - \mu_s(x) \right) \mathrm{d}x \right| \\
&\leq \left(\int_{\mathbb{R}^d} (f(x) - f^\star)^2 \mu_s(x) \mathrm{d}x \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \left(\rho_s(t, x) - \mu_s(x) \right)^2 \mu_s^{-1} \mathrm{d}x \right)^{\frac{1}{2}} \\
&\leq C(s) \mathrm{e}^{-\lambda_s t} \left\| \rho - \mu_s \right\|_{\mu_s^{-1}},
\end{aligned}$$

where the first inequality applies the Cauchy-Schwarz inequality and

$$C(s) = \left(\int_{\mathbb{R}^d} (f - f^\star)^2 \mu_s \mathrm{d}x \right)^{\frac{1}{2}}$$

is an increasing function of s.

■

We conclude this subsection with the proof of Lemma 13, which is well known and can be found in Bakry et al. (2014) for instance.

Proof [Proof of Lemma 13] It follows from (14) that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\rho_s(t,\cdot)-\mu_s\right\|_{\mu_s^{-1}}^2=-s\int_{\mathbb{R}^d}\|\nabla h_s\|^2\mathrm{d}\mu_s.$$

Next, using Lemma 12 and recognizing the equality $\int_{\mathbb{R}^d} h_s\mathrm{d}\mu_s=\int_{\mathbb{R}^d}\rho_s(t,x)\mathrm{d}x=1$, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\rho_s(t,\cdot)-\mu_s\right\|_{\mu_s^{-1}}^2&\leq-2\lambda_s\left(\int_{\mathbb{R}^d}h_s^2\mathrm{d}\mu_s-1\right)\\&=-2\lambda_s\left\|\rho_s(t,\cdot)-\mu_s\right\|_{\mu_s^{-1}}^2.\end{aligned}$$

Integrating both sides yields (17), as desired.

■

5.2 Proof of Proposition 5

Next, we turn to the proof of Proposition 5. We first state a technical lemma, deferring its proof to Appendix C.4.

Lemma 14 *Under the assumptions of Proposition 5, the excess risk at stationarity $\epsilon(s)$ satisfies*

$$\frac{\mathrm{d}\epsilon(0)}{\mathrm{d}s}=0.$$

Using Lemma 14, we now finish the proof of Proposition 5.

Proof [Proof of Proposition 5]

Letting $g=f-f^\star$, we write the excess risk at stationarity as

$$\epsilon(s)=\mathbb{E}f(X_s(\infty))-f^\star=\frac{\int_{\mathbb{R}^d}g\mathrm{e}^{-\frac{2g}{s}}\mathrm{d}x}{\int_{\mathbb{R}^d}\mathrm{e}^{-\frac{2g}{s}}\mathrm{d}x},$$

which yields the following derivative:

$$\frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s}=\frac{\frac{2}{s^2}\int_{\mathbb{R}^d}g^2\mathrm{e}^{-\frac{2g}{s}}\mathrm{d}x\int_{\mathbb{R}^d}\mathrm{e}^{-\frac{2g}{s}}\mathrm{d}x-\frac{2}{s^2}\left(\int_{\mathbb{R}^d}g\mathrm{e}^{-\frac{2g}{s}}\mathrm{d}x\right)^2}{\left(\int_{\mathbb{R}^d}\mathrm{e}^{-\frac{2g}{s}}\mathrm{d}x\right)^2}.$$

Making use of the Cauchy-Schwarz inequality, the derivative satisfies $\frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s}\geq 0$ for all $s>0$. In fact, the equality holds only in the case of a constant f is a constant, which contradicts both the confining condition and the Villani condition. Hence, the inequality can be strengthened to

$$\frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s}>0,$$

for $s>0$. Consequently, we have proven that the excess risk $\epsilon(s)$ at stationarity is a strictly increasing function of $s\in[0,+\infty)$.

Next, from Fatou's lemma we get

$$\epsilon(0) \le \limsup_{s\to 0^+} \epsilon(s) \le \int_{\mathbb{R}^d} \lim_{s\to 0^+} g\mu_s \mathrm{d}x = f^\star - f^\star = 0$$
$$\epsilon(0) \ge \liminf_{s\to 0^+} \epsilon(s) \ge \int_{\mathbb{R}^d} \lim_{s\to 0^+} g\mu_s \mathrm{d}x = f^\star - f^\star = 0.$$

As a consequence, $\epsilon(0) = 0$. Lemma 14 shows that for any $S > 0$, there exists $A = A_S$ such that $0 \le \frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s} \le A$ for all $0 \le s \le S$. This fact, combined with $\epsilon(0) = 0$, immediately gives $\epsilon(s) \le As$ for all $0 \le s \le S$.

■

6. Geometrizing the Exponential Decay Constant

Having established the linear convergence to stationarity for the LR-dependent SDE, we now offer a quantitative characterization of the exponential decay constant λ_s for a class of nonconvex objective functions. This is crucial for us to obtain a clear understanding of the dynamics of SGD and especially its dependence on the learning rate in the nonconvex setting.

6.1 Connection with a Schrödinger operator

We begin by deriving a relationship between the LR-dependent SDE (2) and a Schrödinger operator.[12] Recall that the probability density $\rho_s(t,\cdot)$ of the SDE solution is assumed to be in $L^2(\mu_s^{-1})$. Consider the transformation

$$\psi_s(t,\cdot) = \frac{\rho_s(t,\cdot)}{\sqrt{\mu_s}} \in L^2(\mathbb{R}^d).$$

This transformation allows us to equivalently write the Fokker–Planck–Smoluchowski equation (5) as

$$\frac{\partial \psi_s}{\partial t} = \frac{s}{2}\Delta\psi_s - \left(\frac{\|\nabla f\|^2}{2s} - \frac{\Delta f}{2}\right)\psi_s = -\frac{-s\Delta + V_s}{2}\psi_s, \tag{18}$$

with the initial condition $\psi_s(0,\cdot) = \frac{\rho}{\sqrt{\mu_s}} \in L^2(\mathbb{R}^d)$. $-s\Delta + V_s$ is a Schrödinger perator, where the potential

$$V_s = \frac{\|\nabla f\|^2}{s} - \Delta f$$

is positive for sufficiently large $\|x\|$ due to the Villani condition.

Now, we collect some basic facts concerning the spectrum of the Schrödinger operator $-s\Delta + V_s$. First, it is a positive semidefinite operator, as shown below. Recognizing the uniqueness of the Gibbs distribution (6), it is not hard to show that $\sqrt{\mu_s}$ is the unique

12. The theory of Schrödinger operators is a major component of classical spectral theory; please see the references Hislop and Sigal (2012); Helffer (2013); Reed and Simon (1978).

eigenfunction of $-s\Delta + V_s$ with a corresponding eigenvalue of zero. Using this fact, from the proof of Lemma 13, we get

$$\begin{aligned}
\langle(-s\Delta + V_s)\psi_s(t,\cdot), \psi_s(t,\cdot)\rangle &= \langle(-s\Delta + V_s)(\psi_s(t,\cdot) - \sqrt{\mu_s}), \psi_s(t,\cdot) - \sqrt{\mu_s}\rangle \\
&= -\frac{\mathrm{d}}{\mathrm{d}t}\langle\psi_s(t,\cdot) - \sqrt{\mu_s}, \psi_s(t,\cdot) - \sqrt{\mu_s}\rangle \\
&= -\frac{\mathrm{d}}{\mathrm{d}t}\|\rho_s(t,\cdot) - \mu_s\|^2_{\mu_s^{-1}} \\
&= s\int_{\mathbb{R}^d} \|\nabla(\rho_s(t,\cdot)\mu_s^{-1})\|^2 \mathrm{d}\mu_s \\
&\geq 0,
\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product in $L^2(\mathbb{R}^d)$. In fact, this inequality can be extended to $\langle(-s\Delta + V_s)g, g\rangle \geq 0$ for any g. This verifies the positive semidefiniteness of the Schrödinger operator $-s\Delta + V_s$.

Next, making use of the fact that $\frac{1}{s}V_s(x) \to +\infty$ as $\|x\| \to +\infty$, we state the following well-known result in spectral theory—that the Schrödinger operator has a purely discrete spectrum in $L^2(\mathbb{R}^d)$ Hislop and Sigal (2012). A spectrum is said to be discrete if it takes on distinct eigenvalues, with gaps between one value and the next (see, for example, (Hislop and Sigal, 2012, Definition 1.4)).

Lemma 15 (Theorem 10.7 in Hislop and Sigal (2012)) *Assume that V is continuous, and $V(x) \to +\infty$ as $\|x\| \to +\infty$. Then the operator $-\Delta + V$ has a purely discrete spectrum.*

Taken together, the positive semidefiniteness of $-s\Delta + V_s$ and Lemma 15 allow us to order the eigenvalues of $-s\Delta + V_s$ in $L^2(\mathbb{R}^d)$ as

$$0 = \zeta_{s,0} < \zeta_{s,1} \leq \cdots \leq \zeta_{s,\ell} \leq \cdots < +\infty.$$

Let $\{\psi_{s,i}(x)\}_{i=0}^{\infty}$ represent the eigenfunctions of the Schrödinger operator $-s\Delta + V_s$ in $L^2(\mathbb{R}^d)$. The solution to the equivalent form of the Fokker–Planck–Smoluchowski equation (18) can be expressed in the following form:

$$\psi_s(t,x) = \sum_{i=0}^{\infty} c_i(t)\psi_{s,i}(x). \tag{19}$$

By substituting (19) into (18), we obtain the following equality:

$$\sum_{i=0}^{\infty} \dot{c}_i(t)\psi_{s,i}(x) = -\frac{1}{2}\sum_{i=0}^{\infty} c_i(t)(-s\Delta + V_s)\psi_{s,i}(x) = -\frac{1}{2}\sum_{i=0}^{\infty} \zeta_{s,i}c_i(t)\psi_{s,i}(x).$$

Additionally, we know that the coefficients decay exponentially in t:

$$c_i(t) = e^{-\frac{1}{2}\zeta_{s,i}t}c_i(0).$$

Therefore, the closed-form solution to (18) is

$$\psi_s(t,\cdot) = \sum_{i=0}^{\infty} e^{-\frac{1}{2}\zeta_{s,i}t} c_i(0)\psi_{s,i}(\cdot).$$

A crucial fact from this representation is that the exponential decay constant λ_s in Lemma 13 can be set to

$$\lambda_s = \frac{1}{2}\zeta_{s,1}. \tag{20}$$

To see this, note that $\psi_s(t,\cdot) - \sqrt{\mu_s}$ also satisfies (18) and is orthogonal to the null eigenfunction $\sqrt{\mu_s}$. Therefore, the norm of $\psi_s(t,\cdot) - \sqrt{\mu_s}$ must decay exponentially at a rate determined by half of the smallest positive eigenvalue of H_s.[13] That is, we have

$$\begin{aligned}\langle \psi_s(t,\cdot) - \sqrt{\mu_s}, \psi_s(t,\cdot) - \sqrt{\mu_s}\rangle &\leq \mathrm{e}^{-2\frac{\zeta_{s,1}}{2}t} \langle \psi_s(0,\cdot) - \sqrt{\mu_s}, \psi_s(0,\cdot) - \sqrt{\mu_s}\rangle \\ &= \mathrm{e}^{-\zeta_{s,1}t} \langle \psi_s(0,\cdot) - \sqrt{\mu_s}, \psi_s(0,\cdot) - \sqrt{\mu_s}\rangle,\end{aligned}$$

which is equivalent to

$$\|\rho_s(t,\cdot) - \mu_s\|_{\mu_s^{-1}} \leq \mathrm{e}^{-\frac{\zeta_{s,1}}{2}t} \|\rho - \mu_s\|_{\mu_s^{-1}}.$$

As such, we can take $\lambda_s = \frac{1}{2}\zeta_{s,1}$ in the proof of Lemma 13.

As a consequence of this discussion, we seek to study the Fokker–Planck–Smoluchowski equation (5) by analyzing the spectrum of the linear Schrödinger operator (18), especially its smallest positive eigenvalue $\delta_{s,1}$. To facilitate the analysis, a crucial observation is that this Schrödinger operator is equivalent to the *Witten-Laplacian*,

$$\Delta_f^s := s(-s\Delta + V_s) = -s^2\Delta + \|\nabla f\|^2 - s\Delta f, \tag{21}$$

by a simple scaling. Denoting by the eigenvalues of the Witten-Laplacian as $0 = \delta_{s,0} < \delta_{s,1} \leq \cdots \leq \delta_{s,\ell} \leq \cdots < +\infty$, we obtain the simple relationship

$$\delta_{s,\ell} = s\zeta_{s,\ell},$$

for all ℓ.

The spectrum of the Witten-Laplacian has been the subject of a large literature Helffer and Nier (2005); Bovier et al. (2005); Nier (2004); Arnol'd and Khesin (1999), and in the next subsection, we exploit this literature to derive a closed-from expression for the first positive eigenvalue of the Witten-Laplacian, thereby obtaining the dependence of the exponential decay constant on the learning rate for a certain class of nonconvex objective functions Hérau et al. (2011); Michel (2019).

13. Here, the norm of $\psi_s(t,\cdot) - \sqrt{\mu_s}$ is induced by the inner product in $L^2(\mathbb{R}^d)$. That is,

$$\|\psi(t,\cdot) - \sqrt{\mu_s}\|_{L^2(\mathbb{R}^d)} = \sqrt{\langle \psi(t,\cdot) - \sqrt{\mu_s}, \psi(t,\cdot) - \sqrt{\mu_s}\rangle}.$$

6.2 The spectrum of the Witten-Laplacian: nonconvex Morse functions

We proceed by imposing the mild condition on the objective function that its first-order and second-order derivatives cannot be both degenerate anywhere. Put differently, the objective function is a Morse function. This allows us to use the theory of Morse functions to provide a geometric interpretation of the spectrum of the Witten-Laplacian.

Basics of Morse theory. We give a brief introduction to Morse theory at the minimum level that is necessary for our analysis. Let f be an infinitely differentiable function defined on $\mathbb{R}^n$. A point x is called a critical point if the gradient $\nabla f(x) = 0$. A function f is said to be a Morse function if for any critical point x, the Hessian $\nabla^2 f(x)$ at x is nondegenerate; that is, all the eigenvalues of the Hessian are nonzero. The objective f is assumed to be a Morse function throughout Section 6.2. Note also that we refer to a point x as a local minimum if x is a critical point and all eigenvalues of the Hessian at x are positive.

Next, we define a certain type of saddle point. To this end, let $\eta_1(x) \geq \eta_2(x) \geq \cdots \geq \eta_d(x)$ be the eigenvalues of the Hessian $\nabla^2 f(x)$ at x.[14] A critical point x is said to be an *index-1 saddle point* if the Hessian at x has exactly one negative eigenvalue, that is, $\eta_1(x) \geq \cdots \geq \eta_{d-1}(x) > 0,\ \eta_d(x) < 0$. Of particular importance to this paper is a special kind of index-1 saddle point that will be used to characterize the exponential decay constant. Letting $\mathcal{K}_\nu := \{x \in \mathbb{R}^d : f(x) < \nu\}$ denote the sublevel set at level ν, for an index-1 saddle point x, it is intuitive to imagine that the set $\mathcal{K}_{f(x)} \cap \{x' : \|x' - x\| < r\}$ can be partitioned into two connected components, say $C_1(x, r)$ and $C_2(x, r)$, if the radius r is sufficiently small. The following definition rigorously differentiates index-1 separating saddle points from the other saddle points.

Definition 16 *Let x be an index-1 saddle point and $r > 0$ be sufficiently small. If $C_1(x, r)$ and $C_2(x, r)$ are contained in two different (maximal) connected components of the sublevel set $\mathcal{K}_{f(x)}$, we call x an index-1 separating saddle point.*

The remainder of this section aims to relate index-1 separating saddle points to the convergence rate of the LR-dependent SDE. For ease of reading, the remainder of the paper uses x° to denote an index-1 separating saddle point and writes $\mathcal{X}^\circ$ for the set of all these points. To give a geometric interpretation of Definition 16, let $x_1^\bullet$ and $x_2^\bullet$ denote local minima in the two maximal connected components of $\mathcal{K}_{f(x^\circ)}$, respectively. Intuitively speaking, the index-1 separating saddle point x° is the bottleneck of any path connecting the two local minima. More precisely, along a path connecting $x_1^\bullet$ and $x_2^\bullet$, by definition the function f must attain a value that is at least as large as $f(x^\circ)$. In this regard, the function value at x° plays a fundamental role in determining how long it takes for the LR-dependent SDE initialized at $x_1^\bullet$ to arrive at $x_2^\bullet$. See an illustration in Figure 9.

As is assumed in this section, f is a Morse function and satisfies both the confining and the Villani conditions; in this case, it can be shown that the number of the critical points of f is finite. Thus, denote by n° the number of index-1 separating saddle points of f and let $n^\bullet$ denote the number of local minima.

14. Note that here we order the eigenvalues from the largest to the smallest, as opposed to the case of the Schrödinger operator previously.

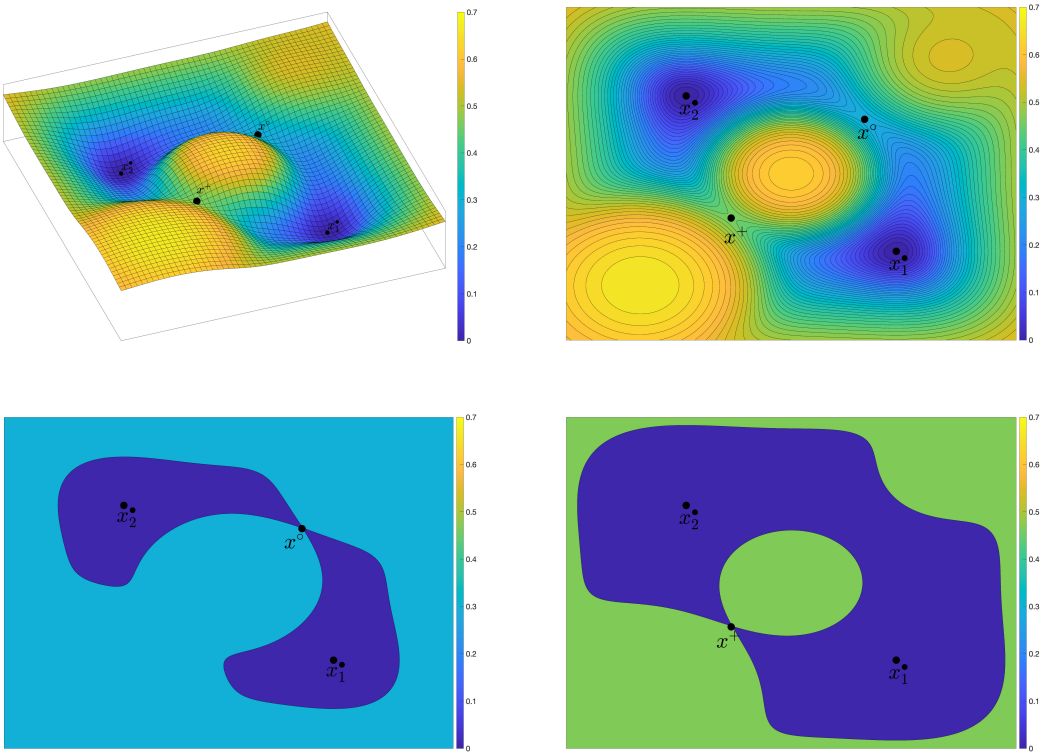

Figure 9: The landscape of a two-dimensional nonconvex Morse function. Here, $x_1^\bullet$ and $x_2^\bullet$ denote two local minima. Both x° and x^+ are index-1 saddle points, but only the former is an index-1 separating saddle point since $f(x^\circ) < f(x^\bullet)$. In the two bottom plots, the deep blue regions form the sublevel sets at $f(x^\circ)$ or $f(x^\bullet)$. Note that the sublevel set induced by x° is the union of two connected components.

Hérau–Hitrik–Sjöstrand's generic case. To describe the labeling procedure, consider the set of the objective values at index-1 separating saddle points $\mathcal{V} = \{f(x^\circ) : x^\circ \in \mathcal{X}^\circ\}$. This is a finite set and we use I to denote the cardinality of this set. Write $\mathcal{V} = \{\nu_1, \ldots, \nu_I\}$ and sort these values as

$$+\infty = \nu_0 > \nu_1 > \cdots > \nu_I, \tag{22}$$

where by convention $\nu_0 = +\infty$ corresponds to a fictive saddle point at infinity.

Next, we follow Hérau et al. (2011) and define a type of connected components of sublevel set.

Definition 17 *A connected component E of the sublevel set $\mathcal{K}_\nu$ for some $\nu \in \mathcal{V}$ is called a critical component if either $\partial E \cap \mathcal{X}^\circ \neq \varnothing$ or $E = \mathbb{R}^d$, where ∂E is the boundary of E.*

In this definition, the case of $E = \mathbb{R}^d$ applies only if $\nu = \nu_0 = +\infty$. If $\nu = \nu_i$ for some $1 \leq i \leq I$ is only attained by one index-1 separating saddle point, the sublevel set $\mathcal{K}_{\nu_i}$ has two critical components. See Definition 16 for more details.

With these preparatory notions in place, we describe the following procedure for labeling index-1 separating saddle points and local minima Hérau et al. (2011). See Figure 10 for an illustration of this process.

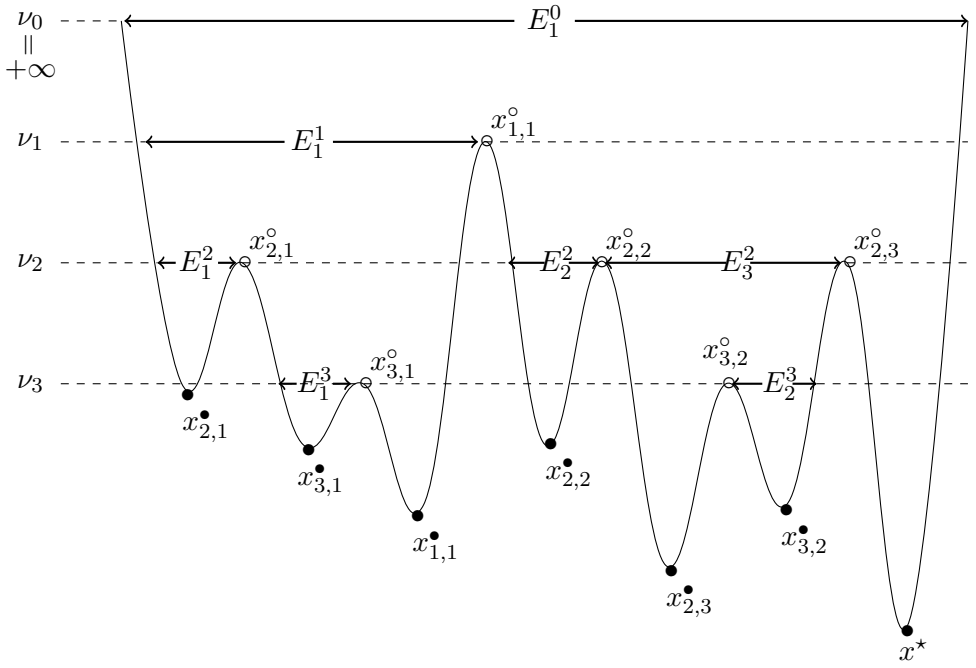

Figure 10: A generic one-dimensional Morse function. The labeling process gives rise to a one-to-one correspondence between the local minimum $x_{ij}^\bullet$ and the index-1 separating saddle point $x_{i,j}^\circ$ (which are also local maxima) for all i, j.

1. Let $E_1^0 := \mathbb{R}^d$. Note that the global minimum $x^\star$ is contained in E_1^0 and denote

$$x_0^\bullet := x^\star = \operatorname*{argmin}_{x \in E_1^0} f(x).$$

   Let $\mathcal{X}_0^\bullet$ denote the singleton set $\{x^\star\}$.

2. Let E_j^1 for $j = 1, \ldots, m_1$ be the critical components of the sublevel set $\mathcal{K}_{\nu_1}$. Note that $E_1^1 \cup \cdots \cup E_{m_1}^1$ is a (proper) subset of $\mathcal{K}_{\nu_1}$. Without loss of generality, assume $x^\star \in E_{m_1}^1$. Then, we select $x_{1,j_1}^\bullet$ as

$$x_{1,j_1}^\bullet = \operatorname*{argmin}_{x \in E_{j_1}^1} f(x).$$

   Define $\mathcal{X}_1^\bullet := \{x_{1,1}^\bullet, \ldots, x_{1,m_1-1}^\bullet\}$.

3. For $i = 2, \ldots, I$, let E_j^i for $j = 1, \ldots, m_i$ be the critical components of the sublevel set $\mathcal{K}_{\nu_i}$. Without loss of generality, we assume that the critical components are ordered

such that there exists an integer $k_i \leq m_i$ satisfying

$$\left(\bigcup_{j=1}^{k_i} E_j^i\right) \bigcap \left(\bigcup_{\ell=0}^{i-1} \mathcal{X}_\ell^\bullet\right) = \varnothing$$

and

$$E_j^i \bigcap \left(\bigcup_{\ell=0}^{i-1} \mathcal{X}_\ell^\bullet\right) \neq \varnothing,$$

for any $j = k_i + 1, \ldots, m_i$. Set $x_{i,j}^\bullet$ to

$$x_{i,j}^\bullet = \operatorname*{argmin}_{x \in E_j^i} f(x),$$

for $j = 1, \ldots, k_i$. Define $\mathcal{X}_i^\bullet := \{x_{i,1}^\bullet, \ldots, x_{i,k_i}^\bullet\}$.

To make the labeling process above valid, however, we need to impose the following assumption on the objective. This assumption is generic in the sense that it should be satisfied by a *generic* Morse function.

Assumption 18 (Generic case Hérau et al. (2011)) *For every critical component E_j^i selected in the labeling process above, where $i = 0, 1, \ldots, I$, we assume that*

- *The minimum $x_{i,j}^\bullet$ of f in any critical component E_j^i is unique.*
- *If $E_j^i \cap \mathcal{X}^\circ \neq \varnothing$, there exists a unique $x_{i,j}^\circ \in E_j^i \cap \mathcal{X}^\circ$ such that $f(x_{i,j}^\circ) = \max_{x \in E_j^i \cap \mathcal{X}^\circ} f(x)$. In particular, $E_j^i \cap \mathcal{K}_{f(x_{i,j}^\circ)}$ is the union of two distinct critical components.*

The first condition in this assumption requires that there exists a unique minimum of the objective f in every critical component E_j^i. In particular, the global minimum $x^\star$ is unique under this assumption. In addition, the second condition requires that among all index-1 separating saddle points in E_j^i, if any, f attains the maximum at exactly one of these points.

Under Assumption 18, the above labeling process includes all the local minima of f. Moreover, it reveals a remarkable result: there exists a bijection between the set of local minima and the set of index-1 separating saddle points (including the fictive one) $\mathcal{X}^\circ \cup \{\infty\}$. As shown in the labeling process, for any local minimum $x_{i,j}^\bullet$, we can relate it to the index-1 separating saddle point at which f attains the maximum in the critical component E_j^i. See Figure 10 for an illustrative example. Interestingly, this shows that the number of local minima is always larger than the number of index-1 separating saddle points by one; that is, $n^\circ = n^\bullet - 1$.

In light of these facts, we can relabel the index-1 separating saddle points x_ℓ° for $\ell = 0, 1, \ldots, n^\circ$ with $x_0^\circ = \infty$, and the local minima $x_\ell^\bullet$ for $\ell = 0, 1, \ldots, n^\bullet - 1$ with $x_0^\bullet = x^\star$, such that

$$f(x_0^\circ) - f(x_0^\bullet) > f(x_1^\circ) - f(x_1^\bullet) \geq \ldots \geq f(x_{n^\bullet-1}^\circ) - f(x_{n^\bullet-1}^\bullet), \tag{23}$$

where $f(x_0^\circ) - f(x_0^\bullet) = f(\infty) - f(x^\star) = +\infty$. A detailed description of this bijection is given in (Hérau et al., 2011, Proposition 5.2).

With the pairs $(x_\ell^\circ, x_\ell^\bullet)$ in place, we readily state the following fundamental result concerning the first $n^\bullet - 1$ smallest positive eigenvalues of the Witten-Laplacian Δ_f^s in (21). Recall that the nonconvex Morse function f satisfies the confining condition and the Villani condition.

Proposition 19 (Theorem 1.2 in Hérau et al. (2011)) *Under Assumption 18 and the assumptions of Theorem 2, there exists $s_0 > 0$ such that for any $s \in (0, s_0]$, the first $n^\bullet - 1$ smallest positive eigenvalues of the Witten-Laplacian Δ_f^s associated with f satisfy*

$$\delta_{s,\ell} = s\left(\gamma_\ell + o(s)\right) \mathrm{e}^{-\frac{2(f(x_\ell^\circ) - f(x_\ell^\bullet))}{s}}$$

for $\ell = 1, 1, \ldots, n^\bullet - 1$, where

$$\gamma_\ell = \frac{|\eta_d(x_\ell^\circ)|}{\pi} \left(\frac{\det(\nabla^2 f(x_\ell^\bullet))}{-\det(\nabla^2 f(x_\ell^\circ))} \right)^{\frac{1}{2}}, \tag{24}$$

and $\eta_d(x_\ell^\circ)$ is the unique negative eigenvalue of $\nabla^2 f(x_\ell^\circ)$.

Using Proposition 19 in conjunction with the simple relationship between the exponential decay constant and the spectrum of the Schrödinger operator/Witten-Laplacian (20), it is a stone's throw to prove Theorem 2 when f is generic. First, we give the definition of the *Morse saddle barrier*.

Definition 20 *Let f satisfy the assumptions of Theorem 2. We call $H_f = f(x_1^\circ) - f(x_1^\bullet)$ the* Morse saddle barrier *of f.*

Proof [Proof of Theorem 2 in the generic case] By Proposition 19, we can set the exponential decay constant to

$$\lambda_s = \frac{1}{2s}\delta_{s,1} = \left(\frac{|\eta_d(x_1^\circ)|}{2\pi} \left(\frac{\det(\nabla^2 f(x_1^\bullet))}{-\det(\nabla^2 f(x_1^\circ))} \right)^{\frac{1}{2}} + o(s) \right) \mathrm{e}^{-\frac{2H_f}{s}}$$

in Theorem 2. Taking $\alpha = \frac{1}{2}\frac{|\eta_d(x_1^\circ)|}{2\pi} \left(\frac{\det(\nabla^2 f(x_1^\bullet))}{-\det(\nabla^2 f(x_1^\circ))} \right)^{\frac{1}{2}}$ in (9), we complete the proof when f falls into the generic case.

■

However, the generic assumption for the labeling process is complex, leading to the lack of a geometric interpretation of the objective function required for the labeling process. To gain further insight, we present a simplifying assumption that is a special case of Assumption 18. This simplification is due to Nier (2004).

Assumption 21 (Simplified generic case Nier (2004)) *The objective functions f takes different values at its local minima and index-1 separating saddle points. That is, letting x_1 be a local minimum or an index-1 separating saddle point, and x_2 likewise, then $f(x_1) \neq f(x_2)$. Furthermore, the differences $f(x_{\ell_1}^\circ) - f(x_{\ell_2}^\bullet)$ are distinct for any ℓ_1 and ℓ_2.*

The following result follows immediately from Proposition 19.

Corollary 22 (Theorem 3.1 in Nier (2004)) *Under Assumption 21 and the assumptions of Theorem 2, Proposition 19 holds. Therefore, Theorem 2 holds in this case.*

Michel's degenerate case. We say that a Morse function is *degenerate* if it satisfies the assumptions of Theorem 2 but not Assumption 18. To violate the generic assumption, for example, we can change the objective value $f(x^{\bullet}_{3,1})$ to $f(x^{\bullet}_{1,1})$ or change $f(x^{\bullet}_{3,2})$ to $f(x^{\bullet}_{2,3})$ in Figure 10. In this situation, the first condition in Assumption 18 is not satisfied. Alternatively, if the objective value at $x^{\circ}_{3,1}$ is changed to $f(x^{\circ}_{2,1})$, the second condition in Assumption 18 is not met. Figure 11 presents an example of a degenerate Morse function.

The main challenge in the degenerate case is the lack of uniqueness of the pairs $(x^{\circ}_{\ell}, x^{\bullet}_{\ell})$ derived from the labeling process. Nevertheless, the uniqueness can be maintained if we work on the function values. Explicitly, the labeling process can be adapted to the degenerate case and still yields unique pairs $(f(x^{\circ}_{\ell}), f(x^{\bullet}_{\ell}))$ obeying

$$f(\infty) - f(x^\star) = f(x^{\circ}_0) - f(x^{\bullet}_0) > f(x^{\circ}_1) - f(x^{\bullet}_1) \geq \ldots \geq f(x^{\circ}_{n^\bullet - 1}) - f(x^{\bullet}_{n^\bullet - 1}).$$

In particular, the number of local minima remains larger than that of index-1 separating saddle points by one in this case. The following result extends Proposition 19 to the degenerate case, which is adapted from Theorem 2.8 of Michel (2019).

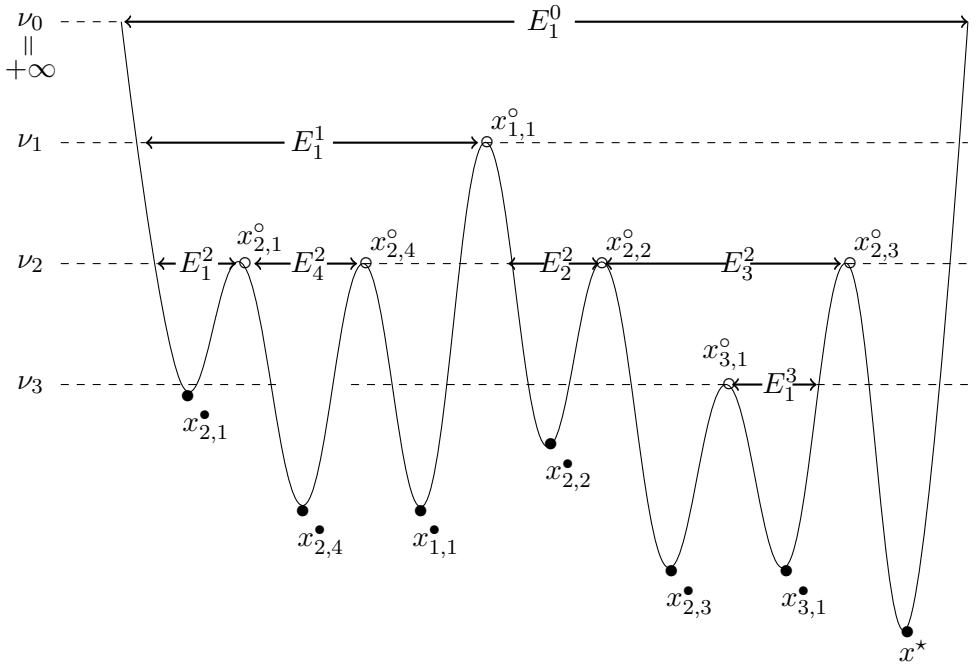

Figure 11: A degenerate one-dimensional Morse function. The labeling of its index-1 separating saddle points $x^{\circ}_{i,j}$ and local minima $x^{\bullet}_{i,j}$ is not unique. Nevertheless, the labeling process gives a unique one-to-one correspondence between the function values at the two types of points. See Figure 10 for a comparison.

Proposition 23 (Theorem 2.8 in Michel (2019)) *Assume that the assumptions of Theorem 2 are satisfied but not Assumption 18. Then, there exists $s_0 > 0$ such that for any*

$s \in (0, s_0]$, the first $n^\bullet - 1$ smallest positive eigenvalues of the Witten-Laplacian Δ_f^s associated with f satisfy

$$\delta_{s,\ell} = s\left(\gamma_\ell + o(s)\right) \mathrm{e}^{-\frac{2H_{f,\ell}}{s}},$$

for $\ell = 1, \ldots, n^\bullet - 1$, where $f(x_\ell^\circ) - f(x_\ell^\bullet) \leq H_{f,\ell} \leq f(x_1^\circ) - f(x^\star)$. The constants $H_{f,\ell}$ and γ_ℓ all depend only on the function f.

Taken together, Proposition 19 and Proposition 23 yield a full proof of Theorem 2. As is clear, the Morse saddle barrier in Definition 20 for the degenerate case is set to $H_f = H_{f,1}$. For completeness, we remark that this result applies to Assumption 18, in which case we conclude that $H_{f,\ell} = f(x_\ell^\circ) - f(x_\ell^\bullet)$ and γ_ℓ is given the same as (24). As such, Proposition 19 is implied by Proposition 23.

7. Discussion

In this paper, we have presented a theoretical perspective on the convergence of SGD in nonconvex optimization as a function of the learning rate. Introducing the notion of an LR-dependent SDE, we have leveraged advanced tools from the study of diffusions, in particular the spectral theory of diffusion operators, to analyze the dynamics of SGD in a continuous-time model. Our findings demonstare that, under certain regularity conditions, the solution to the SDE converges linearly to stationarity. Additionally, we have presented a concise expression for the linear rate of convergence, which transparently depend on the learning rate for nonconvex Morse functions. Our results show that the linear rate is a constant in the strongly convex case, whereas it decreases rapidly as the learning rate decreases in the nonconvex setting. We have thus uncovered a fundamental distinction between convex and nonconvex problems. As one implication, we note that noise in the gradients plays a more determinative role in stochastic optimization with nonconvex objectives as opposed to convex objectives. We also note that our results provide a justification for the use of a large initial learning rate in training neural networks.

We suggest several avenues for future research to enhance and extend the framework for analyzing stochastic optimization methods via SDEs. One area of particular interest is to explore optimization problems where the objective is not L-smooth.[15] It would be intriguing to extend convex quadratic optimization to the infinite-dimensional case, which involves unbounded linear operators. Such an extension would provide valuable insights into the convergence behaviors of the discrete SGD based on the dynamics of the LR-dependent SDE. It is worth noting that in the finite-dimensional case, gradient descent converges linearly to the convex quadratic function. This result is derived by taking the continuous gradient flow as a perspective rather than relying on the error estimate from the numerical method (Proposition 8). With this in mind, it is reasonable to ask whether Theorem 3 can be improved to

$$\mathbb{E} f(x_k) - f^\star \leq \mathrm{O}(s + (1 - \lambda_s s)^k).$$

It is important to highlight that for any learning rate s, there exists some $\tau > 0$ such that when $\lambda > \tau$, the sequence $\{(1-\lambda s)^k\}_{k=0}^\infty$ diverges as k increases. This divergence exhibits a different iteration behavior compared to $\mathrm{e}^{-\lambda t}$, so this direct analogy may not hold. However,

15. For a rigorous definition of L-smooth objective functions, please refer to (Shi et al., 2022, Section 1.4).

in the linear case, considering the implicit discretization still works, it is possible to obtain the following bound in the infinite-dimensional case:

$$\mathbb{E}f(x_k) - f^\star \leq \mathrm{O}(s + (1 + \lambda_s s)^{-k}).$$

More generally, it would be of interest to extend our results to SDEs with variable-dependence noise variance Dieuleveut et al. (2017); Chaudhari and Soatto (2018); Li et al. (2019a). To widen the scope of this framework, it is important to extend our results to the setting where the gradient noise is heavy-tailed Simsekli et al. (2019). Additionally, from a different angle, it is noteworthy that $(s/2)\Delta\rho_s$ in the Fokker–Planck–Smoluchowski equation (5) corresponds to vanishing viscosity in fluid mechanics. Appendix B.3 presents several open problems from this viewpoint.

Another potential direction that go beyond the scope of the L-smooth condition is to explore the optimization problems involving finite-dimensional objective functions with stronger nonlinearity such as the quartic function $f = \|x\|^4$. It is worth noting that while the continuous gradient flow always converges, we cannot guarantee the convergence of the discrete gradient descent from arbitrary initial $x_0 \in \mathbb{R}^d$. From a different perspective, we can view the quartic function $f = \|x\|^4$ as $\|\nabla^2 f(x)\|_2 \leq L_0 + L_1\|x\|^2$ with $L_0 = 0$ and $L_1 = 12$, which can be seen as another natural generalization of L-smooth objective functions. Behind the remarkable success of deep learning in the industry, the variants of SGD widely used in practice are the Ada-series, including Adagrad Duchi et al. (2011), Adam Tieleman and Hinton (2012) and RMSProp Kingma and Ba (2014). Let us take the Adagrad as a representative example. In the deterministic case, the Adagrad can be written as follows:

$$x_{k+1} = x_k - \frac{s\nabla f(x_k)}{\sqrt{\epsilon + \sum_{i=0}^{k} \|\nabla f(x_i)\|^2}}. \tag{25}$$

With the above generalized L-smooth condition, it is not hard to show the average of iterates of the Adagrad (25) converges as

$$f\left(\frac{x_0 + \cdots + x_{k-1}}{k}\right) \leq \mathrm{O}\left(\frac{1}{k}\right).$$

Furthermore, by performing some basic transformations, we can rewrite this equation (25) as

$$\frac{x_{k+1} - x_k}{s^2} = -\frac{\nabla f(x_k)}{\sqrt{\epsilon s^2 + \sum_{i=0}^{k} s^2\|\nabla f(x_i)\|^2}}.$$

Then, by taking the lowest-order continuous limit, we obtain the Adagrad flow as

$$\dot{X} = -\frac{\nabla f(X)}{\sqrt{\int_0^t \|\nabla f(X)\|^2 ds}}. \tag{26}$$

By considering the error, $f(X) - f(x^\star)$, as a Lyapunov function, it becomes evident that the error decreases in (26). Furthermore, it is also possible to explore the evolution of the

probability density by introducing noise to the nonconvex objective function. In practical terms, it seems promising to extend our SDE-based analysis to various learning rate schedules used in practice in training deep neural networks, such as diminishing learning rate and cyclical learning rates Bottou et al. (2018); Smith (2017).

We note also that our results could be useful in guiding the choice of hyperparameters of deep neural networks from an optimization viewpoint. For instance, recognizing the essence of the exponential decay constant λ_s in determining the convergence rate of SGD, it is of interest to consider how to choose the neural network architecture and the loss function so as to get a small value of the Morse saddle barrier H_f. Indeed, Nelson (1966) proposed a stochastic interpretation of quantum mechanics, indicating that the non-deterministic nature of quantum particles could be explained by a stochastic process similar to Brownian motion in classical mechanics. Moreover, the concept of stochastic quantization is introduced in (Parisi and Wu, 1981) to simulate the classical field theory. Currently, based on the reverse viewpoint of stochastic quantization, the quantum algorithm has been explored to speed up the computation compared to local update Metropolis sampling as the ratio of the barrier height over the temperature ratio increases (Mazzola, 2021). Finally, we wonder if the LR-dependent SDE might give insights into generalization properties of neural networks such as local elasticity He and Su (2020) and implicit regularization Zhang et al. (2016); Gunasekar et al. (2018).

Acknowledgments

We would like to thank Zhuang Liu and Yu Sun for helpful conversations about the practical aspects of deep learning. We thank the action editor and three referees for their constructive comments that helped improve the presentation of this work. Bin Shi was supported by grant no.YSBR-034 of CAS. This work was also supported in part by NSF through CAREER DMS-1847415, CCF-1763314, CCF-1934876, and the Wharton Dean's Research Fund. In addition, we recognize support from the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

V. Arnol'd. *Geometrical Methods in the Theory of Ordinary Differential Equations*. Springer Science & Business Media, 2012.

V. Arnol'd. *Mathematical Methods of Classical Mechanics*. Springer Science & Business Media, 2013.

V. I. Arnol'd and B. A. Khesin. *Topological Methods in Hydrodynamics*, volume 125. Springer Science & Business Media, 1999.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer Science & Business Media, 2013.

Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60 – 66, 2008.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, Cham, 2014.

V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations: Ii. convergence rate of the density. *Monte Carlo Methods and Applications*, 2(2):93–128, 1996.

Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer, 2012.

Vladimir I Bogachev, Nikolai V Krylov, and Michael Röckner. Elliptic and parabolic equations for measures. *Russian Mathematical Surveys*, 64(6):973, 2009.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

K. Caluya and A. Halder. Gradient flow algorithms for density propagation in stochastic systems. *IEEE Transactions on Automatic Control*, 2019.

P. Cannarsa and C. Sinestrari. *Semiconcave Functions, Hamilton-Jacobi Equations, and Optimal Control*. Springer Science & Business Media, 2004.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, 2018.

G.-Q. Chen and H. Frid. Vanishing viscosity limit for initial-boundary value problems for conservation laws. *Contemporary Mathematics*, 238:35–51, 1999.

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

A. Chorin and J. Marsden. *A Mathematical Introduction to Fluid Mechanics*. Springer, 1990.

M. Crandall and P.-L. Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

M. Crandall, L. Evans, and P.-L. Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 282(2):487–502, 1984.

D. Davis, D. Drusvyatskiy, and V. Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

J.-D. Deuschel and D. Stroock. *Large deviations*, volume 342. American Mathematical Society, 2001.

J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: A Hamiltonian perspective. *arXiv preprint arXiv:1906.00436*, 2019.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *arXiv preprint arXiv:1707.06386*, 2017.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Andreas Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166:851–886, 2016.

L. Evans. On solving certain nonlinear partial differential equations by accretive operator methods. *Israel Journal of Mathematics*, 36(3-4):225–247, 1980.

L. Evans. *Partial Differential Equations (Second Edition)*, volume 19. American Mathematical Society, 2010.

L. Evans. *An Introduction to Stochastic Differential Equations*, volume 82. American Mathematical Society, 2012.

M. Freidlin and A. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260. Springer Science & Business Media, 2012.

S. Gasiorowicz. *Quantum Physics*. John Wiley & Sons, 2007.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

E. Hazan, A. Rakhlin, and P. Bartlett. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems*, pages 65–72, 2008.

H. He and W. J. Su. The local elasticity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

B. Helffer and F. Nier. *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*. Springer, 2005.

B. Helffer, M. Klein, and F. Nier. Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Mat. Contemp.*, 26:41–85, 2004.

Bernard Helffer. *Spectral theory and its applications*. Number 139. Cambridge University Press, 2013.

F. Hérau, M. Hitrik, and J. Sjöstrand. Tunnel effect and symmetries for Kramers–Fokker–Planck type operators. *Journal of the Institute of Mathematics of Jussieu*, 10(3):567–634, 2011.

Morris W Hirsch, Stephen Smale, and Robert L Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2012.

P. Hislop and I. Sigal. *Introduction to Spectral Theory: With Applications to Schrödinger Operators*, volume 113. Springer Science & Business Media, 2012.

C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

S. Jastrzebski, Z. Kenton, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. *arXiv preprint arXiv:1807.05031*, 2018.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1724–1732. JMLR. org, 2017.

M. I. Jordan. Dynamical, symplectic and stochastic perspectives on gradient-based optimization. In *Proceedings of the International Congress of Mathematicians, Rio de Janeiro*, volume 1, pages 523–550, 2018.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. Tang, and P. Tak. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

P. E. Kloeden and E. Platen. The approximation of multiple stochastic integrals. *Stochastic Analysis and Applications*, 10(4):431–441, 1992.

W. Krichene and P. L. Bartlett. Acceleration and averaging in stochastic descent dynamics. In *Advances in Neural Information Processing Systems*, pages 6796–6806, 2017.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

P. Kundu, I. Cohen, and D. Dowling. *Fluid Mechanics (Fourth Edition)*. Elsevier, 2008.

H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Science & Business Media, 2003.

H. J. Kushner. Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo. *SIAM Journal on Applied Mathematics*, 47(1):169–185, 1987.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2101–2110. JMLR. org, 2017.

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20 (1):1474–1520, 2019a.

Y. Li, C. Wei, and T. Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11669–11680, 2019b.

Z. Li and S. Arora. An exponential learning rate schedule for deep learning. *arXiv preprint arXiv:1910.07454*, 2019.

P.-L. Lions. *Generalized Solutions of Hamilton-Jacobi Equations*, volume 69. London Pitman, 1982.

I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 354–363, 2016.

P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*, pages 1–29, 1999.

Guglielmo Mazzola. Sampling, rates, and reaction currents through reverse stochastic quantization on quantum computers. *Physical Review A*, 104(2):022431, 2021.

L. Michel. About small eigenvalues of Witten Laplacian. *Pure and Applied Analysis*, 1(2), 2019.

G. N. Mil'shtein. Approximate integration of stochastic differential equations. *Theory of Probability & Its Applications*, 19(3):557–562, 1975.

G. N. Mil'shtein. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986.

Edward Nelson. Derivation of the schrödinger equation from newtonian mechanics. *Physical Review*, 150(4):1079, 1966.

F. Nier. Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Journées Equations aux Dérivées Partielles*, pages 1–17, 2004.

E. Pardoux and D. Talay. Discretization and simulation of stochastic differential equations. *Acta Applicandae Math*, 3:23–47, 1985.

G. Parisi and Y. S. Wu. Perturbation theory without gauge fixing. *Scientia. Sinica*, 24(4): 483–496, 1981.

G. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60. Springer, 2014.

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

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics IV: Analysis of Operators*, volume 4. Elsevier, 1978.

B. Shi, S. Du, M. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195 (1-2):79–148, 2022.

U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*, 2019.

L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

M. Sordello and W. J. Su. Robust learning rate selection for stochastic optimization via splitting diagnostic. *arXiv preprint arXiv:1910.08597*, 2019.

W. J. Su, S. Boyd, and E. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

R. Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

D. Talay. *Analyse numérique des équations différentielles stochastiques*. PhD thesis, Université Aix-Marseille I, 1982.

D. Talay. Efficient numerical schemes for the approximation of expectations of functionals of the solution of a SDE and applications. In *Filtering and Control of Random Processes*, pages 294–313. Springer, 1984.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

C. Villani. Hypocoercive diffusion operators. In *International Congress of Mathematicians*, volume 3, pages 473–498, 2006.

C. Villani. Hypocoercivity. *Memoirs of the American Mathematical Society*, 202(950), 2009.

Martin J Wainwright and Michael I Jordan. A variational principle for graphical models. *New Directions in Statistical Signal Processing*, page 155, 2006.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.

K. You, M. Long, J. Wang, and M. I. Jordan. How does learning rate decay help modern neural networks? *arXiv preprint arXiv:1908.01878*, 2019.

M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Y. Zhang, P. Liang, and M. Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.

Appendix A. Technical Details for Sections 1 and 2

A.1 Approximating differential equations

Figure 12 presents a diagram that shows approximating surrogates for GD, SGD, and SGLD at multiple scales. In the case of SGD, for example, the inclusion of only $O(1)$ terms leads to the ODE $\dot{X} = -\nabla f(X)$, whereas the inclusion of up to $O(\sqrt{s})$ terms leads to the LR-dependent SDE (2). For GD and SGLD, $O(\sqrt{s})$ terms are not found in the expansion as in the derivation of (2). The $O(\sqrt{s})$-approximation, therefore, leads to the same differential equation as the $O(1)$-approximation for both GD and SGLD.

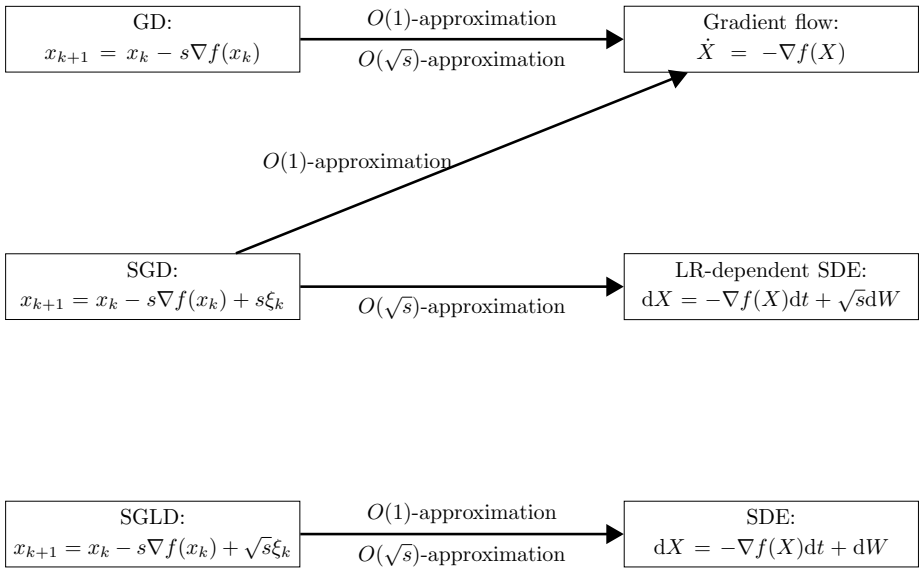

Figure 12: Diagram showing the relationship between three discrete algorithms and their $O(1)$-approximating and $O(1) + O(\sqrt{s})$-approximating differential equations. Note that the inclusion of only $O(1)$-terms does not distinguish between GD and SGD.

A.2 Derivation of the Fokker–Planck–Smoluchowski equation

To derive the LR-dependent Fokker–Planck–Smoluchowski equation (5), we first state the following lemma.

Lemma 24 (Itô's lemma) *For any $f \in C^\infty(\mathbb{R}^d)$ and $g \in C^\infty([0,+\infty)\times\mathbb{R}^d)$, let $X_s(t)$ be the solution to the LR-dependent SDE* (2)*. Then, we have*

$$\mathrm{d}g(t, X_s(t)) = \left(\frac{\partial g}{\partial t} - \nabla f \cdot \nabla g + \frac{s}{2}\Delta g\right)\mathrm{d}t + \sqrt{s}\left(\sum_{i=1}^{d}\frac{\partial g}{\partial x_i}\right)\mathrm{d}W. \tag{27}$$

From this lemma, we get

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}[g(t, X_s(t))|X_s(t')]}{\mathrm{d}t} = \frac{\partial \mathbb{E}[g(t, X_s(t))|X_s(t')]}{\partial t} - \nabla f \cdot \nabla \mathbb{E}[g(t, X_s(t))|X_s(t')] \\ + \frac{s}{2}\Delta \mathbb{E}[g(t, X_s(t))|X_s(t')],\end{aligned} \tag{28}$$

for $t \geq t'$. Setting $v_s(t', x) = \mathbb{E}[g(t, X_s(t))|X_s(t') = x]$. Since $\mathbb{E}[g(t, X_s(t))|X_s(t') = x]$ is invariant with time t, from (28) we see that $v_s(t', x)$ satisfies the following differential equation:

$$\frac{\partial v_s}{\partial t'} = \nabla f \cdot \nabla v_s - \frac{s}{2}\Delta v_s, \qquad v_s(t, x) = g(t, x). \tag{29}$$

Recognizing the invariance of translation of time and letting $u_s(t-t',x) = v_s(t',x)$, we can reduce (29) to the following backward Fokker–Planck–Smoluchowski equation:

$$\frac{\partial u_s}{\partial t} = -\nabla f \cdot \nabla u_s + \frac{s}{2}\Delta u_s, \qquad u_s(0,x) = g(t,x). \tag{30}$$

Next, from the Chapman–Kolmogorov equation, we get

$$\rho_s(t,x) = \int_{\mathbb{R}^d} \rho_s(t,x|0,y)\rho_s(0,y)\mathrm{d}y,$$

where $\rho_s(t,x|0,y) = \rho_s(X(t)=x|X(0)=y)$ and by switching the order of the integration, we obtain

$$\begin{aligned}\int_{\mathbb{R}^d} u_s(0,x)\rho_s(t,x)\mathrm{d}x &= \int_{\mathbb{R}^d} g(t,x)\rho_s(t,x)\mathrm{d}x \\ &= \int_{\mathbb{R}^d} g(t,x)\left(\int_{\mathbb{R}^d} \rho_s(t,x|0,y)\rho_s(0,y)\mathrm{d}y\right)\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \rho_s(0,y)u_s(t,y)\mathrm{d}y = \int_{\mathbb{R}^d} \rho_s(0,x)u_s(t,x)\mathrm{d}x. \end{aligned} \tag{31}$$

Making use of the backward Fokker–Planck–Smoluchowski equation (30) and switching the order of integration (31), we get

$$\begin{aligned}\int_{\mathbb{R}^d} u_s(0,x)\frac{\partial \rho_s(t,x)}{\partial t}\bigg|_{t=0}\mathrm{d}x &= \int_{\mathbb{R}^d} \frac{\partial u_s(t,x)}{\partial t}\bigg|_{t=0}\rho_s(0,x)\mathrm{d}x \\ &= \int_{\mathbb{R}^d} u_s(0,x)\left(\nabla\cdot(\rho_s(0,x)\nabla f(x)) + \frac{s}{2}\Delta\rho_s(0,x)\right)\mathrm{d}x.\end{aligned}$$

Hence, we derive the forward Fokker–Planck–Smoluchowski equation at $t=0$ for an arbitrary smooth function $u_s(0,x) = g(t,x)$. Noting that $t=0$ can be replaced by any time t, we complete the derivation of the Fokker–Planck–Smoluchowski equation.

A.3 The uniqueness of Gibbs invariant distribution

We begin by proving that the probability density μ_s is an invariant distribution of (5). Plugging

$$\nabla\mu_s = -\frac{2}{s}(\nabla f)\,\mu_s$$

into (5) gives

$$\nabla\cdot(\mu_s\nabla f) = \nabla\mu_s\cdot\nabla f + \mu_s\Delta f = -\frac{2}{s}\|\nabla f\|^2\mu_s + (\Delta f)\mu_s \tag{32}$$

and

$$\Delta\mu_s = -\frac{2}{s}\nabla f\cdot\nabla\mu_s - \frac{2}{s}\mu_s\Delta f = \frac{4}{s^2}\|\nabla f\|^2\mu_s - \frac{2}{s}\mu_s\Delta f. \tag{33}$$

Combining (32) and (33) yields

$$\nabla\cdot(\mu_s\nabla f) + \frac{s}{2}\Delta\mu_s = 0.$$

We now proceed to show that the probability density μ_s is unique. To derive a contradiction, we assume that there exists another distribution ϑ_s satisfying the Fokker–Planck–Smoluchowski equation:

$$\nabla \cdot (\vartheta_s \nabla f) + \frac{s}{2} \Delta \vartheta_s = 0. \tag{34}$$

Write $\varpi_s = \vartheta_s \mu_s^{-1}$ and recall the operator $\mathscr{L}_s$ defined in Section 5.1. We can rewrite (34) as

$$\mathscr{L}_s \varpi_s = 0.$$

Using Lemma 11, we have

$$0 = \int_{\mathbb{R}^d} (\mathscr{L}_s \varpi_s) \varpi_s \mathrm{d}\mu_s = -\frac{s}{2} \int_{\mathbb{R}^d} \|\nabla \varpi_s\|^2 \mathrm{d}\mu_s \leq 0.$$

Hence, ϖ_s must be a constant on $\mathbb{R}^d$. Furthermore, since both μ_s and ϑ_s are probability densities, it must be the case that $\varpi_s \equiv 1$. In other words, ϑ_s is identical to μ_s. The proof is complete.

A.4 Proof of Lemma 2

Recall that Section 6.1 shows that the transition probability density $\rho_s(t, x)$ in $C^1([0, +\infty), L^2(\mu_s^{-1}))$ governed by the Fokker–Planck–Smoluchowski equation (5) is equivalent to the function $\psi_s(t, x)$ in $C^1([0, +\infty), L^2(\mathbb{R}^d))$ governed by (18). Moreover, in Section 6.1, we have shown that the spectrum of the Schrödinger operator $-s\Delta + V_s$ satisfies

$$0 = \zeta_{s,0} < \zeta_{s,1} \leq \cdots \leq \zeta_{s,\ell} \leq \cdots < +\infty.$$

Since $L^2(\mathbb{R}^d)$ is a Hilbert space, there exists a standard orthogonal basis corresponding to the spectrum of $-s\Delta + V_s$:

$$\mu_s = \phi_{s,0},\ \phi_{s,1},\ \ldots,\ \phi_{s,\ell},\ \ldots \in L^2(\mathbb{R}^d).$$

Then, for any initialization $\psi_s(0, x) \in L^2(\mathbb{R}^d)$, there exist constants c_ℓ $(\ell = 1, 2, \ldots)$ such that

$$\psi_s(0, \cdot) = \sqrt{\mu_s} + \sum_{\ell=1}^{+\infty} c_\ell \phi_{s,\ell}.$$

Thus, the solution to the partial differential equation (18) is

$$\psi_s(t, \cdot) = \sqrt{\mu_s} + \sum_{\ell=1}^{+\infty} c_\ell \mathrm{e}^{-\zeta_{s,\ell} t} \phi_{s,\ell}.$$

Recognizing the transformation $\psi_s(t, \cdot) = \rho_s(t, \cdot)/\sqrt{\mu_s}$, we recover

$$\rho_s(t, \cdot) = \mu_s + \sum_{\ell=1}^{+\infty} c_\ell \mathrm{e}^{-\zeta_{s,\ell} t} \phi_{s,\ell} \sqrt{\mu_s}.$$

Note that $\zeta_{s,\ell}$ is positive for $\ell \geq 1$. Thus, the proof is finished.

Appendix B. Technical Details for Section 3

B.1 Proof of Lemma 7

Here, we prove Lemma 7 using the Bakry–Emery theorem, which is a Poincaré-type inequality for μ-strongly convex functions. As a direct consequence of this lemma, the exponential decay constant for strongly convex objectives does not depend on the learning rate s and the ambient dimension d.

Lemma 25 (Bakry–Emery theorem) *Let f be an infinitely differentiable function defined on $\mathbb{R}^d$. If f is μ-strongly convex, then the measure $\mathrm{d}\mu_s$ satisfies the Poincaré-type inequality as in Lemma 12 with $\lambda_s = \mu$; that is, for any smooth function h with a compact support,*

$$\int_{\mathbb{R}^d} h^2 \mathrm{d}\mu_s - \left(\int_{\mathbb{R}^d} h \mathrm{d}\mu_s\right)^2 \leq \frac{s}{2\mu} \int_{\mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s.$$

Lemma 25 serves as the main technical tool in the proof of Lemma 7. Its proof is in Appendix B.1.1. Now, we prove the following result using Lemma 25.

Lemma 26 *Under the same assumptions as in Lemma 7, $\rho_s(t,\cdot)$ converges to the Gibbs distribution μ_s in $L^2(\mu_s^{-1})$ at the rate*

$$\|\rho_s(t,\cdot) - \mu_s\|_{\mu_s^{-1}} \leq \mathrm{e}^{-\mu t} \|\rho_s - \mu_s\|_{\mu_s^{-1}}. \tag{35}$$

Proof [Proof of Lemma 26] It follows from (14) that

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\rho_s(t,\cdot) - \mu_s\|^2_{\mu_s^{-1}} = -s \int_{\mathbb{R}^d} \|\nabla h_s\|^2 \mathrm{d}\mu_s.$$

Next, using Lemma 25 and recognizing the equality $\int_{\mathbb{R}^d} h_s \mathrm{d}\mu_s = \int_{\mathbb{R}^d} \rho_s(t,x)\mathrm{d}x = 1$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\rho_s(t,\cdot) - \mu_s\|^2_{\mu_s^{-1}} \leq -2\mu \int_{\mathbb{R}^d} (h_s - 1)^2 \mathrm{d}\mu_s = -2\mu \|\rho_s(t,\cdot) - \mu_s\|^2_{\mu_s^{-1}}.$$

Integrating both sides yields (35), as desired. ■

Leveraging Lemma 26, we proceed to complete the proof of Lemma 7.

Proof [Proof of Lemma 7] Using Lemma 26, we get

$$\leq C(s)\mathrm{e}^{-\mu t} \|\rho - \mu_s\|_{\mu_s^{-1}},$$

where the first inequality applies the Cauchy-Schwarz inequality and

$$C(s) = \left(\int_{\mathbb{R}^d} (f - f^\star)^2 \mu_s \mathrm{d}x\right)^{\frac{1}{2}}$$

is an increasing function of s. ■

B.1.1 Proof of Lemma 25

We introduce two operators Γ_s and $\Gamma_{s,2}$ that are built on top of the linear operator $\mathscr{L}_s$ defined in (16). For any $g_1, g_2 \in L^2(\mu_s)$, let

$$\Gamma_s(g_1, g_2) = \frac{1}{2}\left[\mathscr{L}_s(g_1 g_2) - g_1 \mathscr{L}_s g_2 - g_2 \mathscr{L}_s g_1\right] \tag{36}$$

and

$$\Gamma_{s,2}(g_1, g_2) = \frac{1}{2}\left[\mathscr{L}_s \Gamma_s(g_1, g_2) - \Gamma_s(g_1, \mathscr{L}_s g_2) - \Gamma_s(g_2, \mathscr{L}_s g_1)\right]. \tag{37}$$

A simple relationship between the two operators is described in the following lemma.

Lemma 27 *Under the same assumptions as in Lemma 25, for any $g \in L^2(\mu_s)$ we have*

$$\Gamma_{s,2}(g, g) \geq \mu \Gamma_s(g, g).$$

Proof [Proof of Lemma 27]
Note that

$$\mathscr{L}_s(g_1 g_2) = -g_1(\nabla f \cdot \nabla g_2) - g_2(\nabla f \cdot \nabla g_1) + \frac{s}{2}\left(g_1 \Delta g_2 + g_2 \Delta g_1 + 2\nabla g_1 \cdot \nabla g_2\right)$$

and

$$g_1 \mathscr{L}_s h_2 = -g_1 \nabla f \cdot \nabla g_2 + \frac{s}{2} g_1 \Delta g_2, \qquad g_2 \mathscr{L}_s g_1 = -g_2 \nabla f \cdot \nabla g_1 + \frac{s}{2} g_2 \Delta g_1.$$

Then, the operator Γ_s must satisfy

$$\Gamma_s(g, g) = \frac{s}{2}\left(\nabla g \cdot \nabla g\right). \tag{38}$$

Next, together with the equality

$$\frac{1}{2}\Delta(\|\nabla g\|^2) = \nabla g \cdot \nabla(\Delta g) + \mathbf{Tr}[(\nabla^2 g)^T (\nabla^2 g)],$$

we obtain that the operator $\Gamma_{s,2}$ satisfies

$$\Gamma_{s,2}(g, g) = \frac{s}{2}(\nabla g)^T \nabla^2 f (\nabla g) + \frac{s^2}{4}\mathbf{Tr}[(\nabla^2 g)^T (\nabla^2 g)], \tag{39}$$

where $\mathbf{Tr}$ is the standard trace of a squared matrix. Recognizing that the objective f is μ-strongly convex, a comparison between (38) and (39) completes the proof.

■

Recall that $h_s(t, \cdot) \in L^2(\mu_s)$ is the solution to the partial differential equation (15), with the initial condition $h_s(0, \cdot) = h$. Define

$$\Lambda_{1,s}(t) = \int_{\mathbb{R}^d} h_s^2(t, \cdot) \mathrm{d}\mu_s. \tag{40}$$

The following lemma considers the derivatives of $\Lambda_{1,s}(t)$.

Lemma 28 *Under the same assumptions as in Lemma 25, we have*

$$\dot{\Lambda}_{1,s}(t) = -2\int_{\mathbb{R}^d} \Gamma_s(h_s, h_s)\mathrm{d}\mu_s, \qquad \ddot{\Lambda}_{1,s}(t) = 4\int_{\mathbb{R}^d} \Gamma_{s,2}(h_s, h_s)\mathrm{d}\mu_s. \tag{41}$$

Proof [Proof of Lemma 28]

Taking together (14) and (38), we have

$$\int_{\mathbb{R}^d} \Gamma_s(h_s, h_s)\mu_s\mathrm{d}\mu_s = -\int_{\mathbb{R}^d} h_s\mathscr{L}_s h_s\mathrm{d}\mu_s.$$

Since $h_s(t,\cdot) \in L^2(\mu_s)$ is the solution to the partial differential equation (15), we get

$$\dot{\Lambda}_{1,s}(t) = 2\int_{\mathbb{R}^d} h_s\mathscr{L}_s h_s\mathrm{d}\mu_s = -2\int_{\mathbb{R}^d} \Gamma_s(h_s, h_s)\mathrm{d}\mu_s.$$

Furthermore, by the definition of $\Gamma_{s,2}$ and integration by parts, we have[16]

$$\int_{\mathbb{R}^d} \Gamma_{s,2}(h_s, h_s)\mathrm{d}\mu_s = \int_{\mathbb{R}^d} (\mathscr{L}_s h_s)^2\mathrm{d}\mu_s.$$

From Lemma 11, we know that the linear operator $\mathscr{L}_s$ is self-adjoint. Then, we obtain the second derivative as

$$\ddot{\Lambda}_{1,s}(t) = 2\int_{\mathbb{R}^d} (\mathscr{L}_s h_s)^2\mathrm{d}\mu_s + 2\int_{\mathbb{R}^d} h_s\mathscr{L}_s^2 h_s\mathrm{d}\mu_s = 4\int_{\mathbb{R}^d} \Gamma_{s,2}(h_s, h_s)\mathrm{d}\mu_s.$$

■

Finally, we complete the proof of Lemma 25.

Proof [Proof of Lemma 25] Using Lemma 27 and Lemma 28, we obtain the following inequality:

$$\ddot{\Lambda}_{1,s}(t) \geq -2\mu\dot{\Lambda}_{1,s}(t). \tag{42}$$

From the definition of $\Lambda_{1,s}(t)$, we have

$$\Lambda_{1,s}(0) - \Lambda_{1,s}(\infty) = \int_{\mathbb{R}^d} h^2\mathrm{d}\mu_s - \left(\int_{\mathbb{R}^d} h\mathrm{d}\mu_s\right)^2,$$

where the second term on the right-hand side follows from Lemma 10 and

$$\int_{\mathbb{R}^d} h\mathrm{d}\mu_s = \int_{\mathbb{R}^d} \rho\mathrm{d}x = 1.$$

By Lemma 10, we get $h_s(\infty,\cdot) \equiv 1$, which together with (41) gives

$$\dot{\Lambda}_{1,s}(0) - \dot{\Lambda}_{1,s}(\infty) = -2\int_{\mathbb{R}^d} \Gamma_s(h, h)\mathrm{d}\mu_s = -s\int_{\mathbb{R}^d} \|\nabla h\|^2\mathrm{d}\mu_s.$$

The final equality follows from (38). Integrating both sides of the inequality (42), we have

$$-2\mu\left(\Lambda_{1,s}(0) - \Lambda_{1,s}(\infty)\right) \leq \dot{\Lambda}_{1,s}(0) - \dot{\Lambda}_{1,s}(\infty),$$

which completes the proof. ■

16. See the calculation in Bakry et al. (2013).

B.2 Proof of Proposition 8

By Lemma 2, let $\rho_s(t,\cdot) \in C^1([0,+\infty), L^2(\mu_s^{-1}))$ denote the unique transition probability density of the solution to the LR-dependent SDE. Taking an expectation, we get

$$\mathbb{E}[X_s(t)] = \int_{\mathbb{R}^d} x\rho_s(t,x)\mathrm{d}x.$$

Hence, the uniqueness has been proved. Using the Cauchy–Schwarz inequality and Lemma 13, we obtain:

$$\begin{aligned}\|\mathbb{E}[X_s(t)]\| &\le \left\|\int_{\mathbb{R}^d} x(\rho_s(t,\cdot)-\mu_s)\mathrm{d}x\right\| + \left\|\int_{\mathbb{R}^d} x\mu_s\mathrm{d}x\right\| \\ &\le \left(\int_{\mathbb{R}^d}\|x\|^2 d\mu_s\right)^{\frac{1}{2}}\left(e^{-\lambda_s t}\left\|\rho-\mu_s\right\|_{\mu_s^{-1}}+1\right) < +\infty,\end{aligned}$$

where the integrability $\int_{\mathbb{R}^d}\|x\|^2\mu_s(x)\mathrm{d}x$ is due to the fact that the objective f satisfies the Villani condition. The existence of a global solution to the LR-dependent SDE (2) is thus established.

For the strong convergence, the LR-dependent SDE (2) corresponds to the Milstein scheme in numerical methods. The original result is obtained by Milstein Mil'shtein (1975) and Talay Talay (1982); Pardoux and Talay (1985), independently. We refer the readers to (Kloeden and Platen, 1992, Theorem 10.3.5 and Theorem 10.6.3), which studies numerical schemes for stochastic differential equation. For the weak convergence, we can obtain numerical errors by using both the Euler-Maruyama scheme and Milstein scheme. The original result is obtained by Milstein Mil'shtein (1986) and Talay Pardoux and Talay (1985); Talay (1984) independently and (Kloeden and Platen, 1992, Theorem 14.5.2) is also a well-known reference. Furthermore, there exists a more accurate estimate of $B(T)$ shown in Bally and Talay (1996). The original proofs in the aforementioned references only assume finite smoothness such as $C^6(\mathbb{R}^d)$ for the objective function.

B.3 Connection with vanishing viscosity

Taking $s = 0$, the zero-viscosity steady-state equation of the Fokker–Planck–Smoluchowski equation (5) reads

$$\nabla\cdot(\mu_0\nabla f) = 0. \tag{43}$$

A solution to this zero-viscosity steady-state equation takes the form

$$\mu_0(x) = \sum_{i=1}^m c_i\delta(x-x_i), \quad \text{with} \quad \sum_{i=1}^m c_i = 1, \tag{44}$$

where x_i's are critical points of the objective f. As is clear, the solution is not unique. However, we have shown previously that the invariant distribution μ_s is unique and converges to

$$\mu_{s\to 0}(x) = \delta(x-x^\star)$$

in the sense of distribution, which is a special case of (44). Clearly, when there exists more than one critical point, $\mu_{s\to 0}(x)$ is different from $\mu_0(x)$ in general. In contrast, $\mu_{s\to 0}(x)$

and $\mu_0(x)$ must be the same for (strictly) convex functions. In light of this comparison, the correspondences between the case $s > 0$ and the case $s = 0$ are fundamentally different in nonconvex and convex problems.

Next, we consider the rate of convergence in the convex setting. Let

$$f(x) = \frac{1}{2}\theta x^2,$$

where $\theta > 0$. Plugging into the Fokker-Planck-Smoluchowski equation (5), we have

$$\begin{cases} \dfrac{\partial \rho_s}{\partial t} = \theta \dfrac{\partial (x\rho_s)}{\partial x} + \dfrac{s}{2}\dfrac{\partial^2 \rho_s}{\partial x^2} \\ \rho(0,\cdot) = \rho \in L^2(\sqrt{s\pi/\theta}\mathrm{e}^{\theta x^2/s})). \end{cases} \tag{45}$$

The solution to (45) is

$$\rho_s(t,x) = \sqrt{\frac{\theta}{\pi s\left(1-\mathrm{e}^{-2\theta t}\right)}} \exp\left[-\frac{\theta}{s}\frac{\left(x - x_0\mathrm{e}^{-\theta t}\right)^2}{1-\mathrm{e}^{-2\theta t}}\right]. \tag{46}$$

For any $\phi(x) \in L^2(\sqrt{s\pi/\theta}\mathrm{e}^{\theta x^2/s})$, we have

$$\begin{aligned} \langle \rho_s, \phi\rangle &= \left\langle \sqrt{\frac{\theta}{\pi s\left(1-\mathrm{e}^{-2\theta t}\right)}} \exp\left[-\frac{\theta}{s}\frac{\left(x - x_0\mathrm{e}^{-\theta t}\right)^2}{1-\mathrm{e}^{-2\theta t}}\right], \phi(x)\right\rangle \\ &= \left\langle \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}, \phi\left(\sqrt{\frac{s(1-\mathrm{e}^{-2\theta t})}{2\theta}}\cdot x + x_0\mathrm{e}^{-\theta t}\right)\right\rangle \to \phi\left(x_0\mathrm{e}^{-\theta t}\right) = \left\langle \delta(x - x_0\mathrm{e}^{-\theta t}), \phi(x)\right\rangle \end{aligned}$$

as $s \to 0$, where $\delta(x - x_0\mathrm{e}^{-\theta t})$ denotes the solution to the following zero-viscosity equation

$$\frac{\partial \rho_0}{\partial t} = \nabla \cdot (\rho_0 \nabla f). \tag{47}$$

Furthermore, using the following inequality

$$\left\| \phi\left(\sqrt{\frac{s(1-\mathrm{e}^{-2\theta t})}{2\theta}}\cdot x + x_0\mathrm{e}^{-\theta t}\right) - \phi\left(x_0\mathrm{e}^{-\theta t}\right)\right\|_\infty \le O\left(\sqrt{s}\right),$$

we get $\langle \rho(t,x), \psi(x)\rangle \to \langle \delta(x - x_0\mathrm{e}^{-\theta t}), \psi(x)\rangle$ at the rate $O\left(\sqrt{s}\right)$ for a test function ψ.

The phenomenon presented above is called *singular perturbation*. It appears in mathematical models of boundary layer phenomena (Chorin and Marsden, 1990, Chapter 2.2, Example 1 and Example 2), WKB theory for Schrödinger equations (Gasiorowicz, 2007, Supplement 4A), KAM theory for circle diffeomorphisms (Arnol'd, 2012, Chapter 2, Section 11) and that for Hamilton systems (Arnol'd, 2013, Appendix 8). Moreover, the singular perturbation phenomenon shows that there exists a fundamental distinction between the $O(1)$-approximating ODE for SGD and the LR-dependent SDE (2). In particular, the learning rate $s \to 0$ in the Fokker–Planck–Smoluchowski equation (5) corresponds to vanishing viscosity. The vanishing viscosity phenomenon was originally observed in fluid mechanics Chorin and Marsden (1990); Kundu et al. (2008), particularly in the degeneration of

the Navier–Stokes equation to the Euler equation Chen and Frid (1999). As a milestone, the vanishing viscosity method has been used to study the Hamilton–Jacobi equation Crandall and Lions (1983); Evans (1980); Crandall et al. (1984). In fact, the Fokker–Planck–Smoluchowski equation (5) and its stationary equation are a form of Hamilton–Jacobi equation with a viscosity term, for which the Hamiltonian is

$$H(x, \rho, \nabla\rho) = \Delta f \rho + \nabla f \cdot \nabla \rho. \tag{48}$$

The Hamiltonian (48) is different from the classical case Lions (1982); Cannarsa and Sinestrari (2004); Evans (2010), which is generally nonlinear in $\nabla\rho$ (cf. Burger's equation). Although the Hamiltonian depends linearly on ρ and $\nabla\rho$, the coefficients depend on Δf and ∇f. Hence, it is not reasonable to apply directly the well-established theory of Hamilton–Jacobi equations Crandall and Lions (1983); Evans (1980); Crandall et al. (1984); Lions (1982); Cannarsa and Sinestrari (2004); Evans (2010) to the Fokker–Planck–Smoluchowski equation (5) and its stationary equation. Furthermore, for the aforementioned example, which proves the $O(\sqrt{s})$ convergence for the Fokker–Planck–Smoluchowski equation with the quadratic potential $f(x) = \frac{\theta}{2}x^2$, is also a viscosity solution to the Hamilton–Jacobi equation Crandall and Lions (1983), since the Hamiltonian (48) for the quadratic potential degenerates to

$$H(x, \rho, \nabla\rho) = 2\mathbf{tr}(A)\rho + 2Ax \cdot \nabla\rho,$$

where $f(x) = x^T Ax$ and A is positive definite and symmetric. Thus, we remark that the general theory of viscosity solutions to Hamilton–Jacobi equations cannot be used directly to prove the theorems in the main body of this paper.

In closing, we present several open problems.

- Consider the stationary solution $\mu_s(x)$ to the Fokker–Planck–Smoluchowski equation (5). For a convex or strongly convex objective f with Lipschitz gradients, can we quantify the rate of convergence? Does the rate of convergence remain $O(\sqrt{s})$?

- Let $T > 0$ be fixed and consider the solution $\rho_s(t, x)$ to the Fokker–Planck–Smoluchowski equation (5) in $[0, T]$. For a convex or strongly convex objective f with Lipschitz gradients, does the solution to the Fokker–Planck–Smoluchowski equation (5) converge to the solution to its zero-viscosity equation (47)? Is the rate of convergence still $O(\sqrt{s})$?

- Consider the solution $\rho_s(t, x)$ to the Fokker–Planck–Smoluchowski equation (5) in $[0, +\infty)$. For a convex or strongly convex objective f with Lipschitz gradients, does the global solution to the Cauchy problem of the Fokker–Planck–Smoluchowski equation (5) converge to the solution of its zero-viscosity equation (47)? Is the rate of convergence still $O(\sqrt{s})$?

Appendix C. Technical Details for Section 5

C.1 Proof of Lemma 9

From the Cauchy–Schwarz inequality, we get

$$\int_{\mathbb{R}^d} |g(x)| \mathrm{d}x \leq \left(\int_{\mathbb{R}^d} g^2(x) \mathrm{e}^{\frac{2f(x)}{s}} \mathrm{d}x \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \mathrm{e}^{-\frac{2f(x)}{s}} \mathrm{d}x \right)^{\frac{1}{2}} < +\infty.$$

This completes the proof.

C.2 Proof of Lemma 11

Recall that the linear operator $\mathscr{L}_s$ in (16) is defined as

$$\mathscr{L}_s = -\nabla f \cdot \nabla + \frac{s}{2} \Delta f.$$

Note that we have

$$\int_{\mathbb{R}^d} (\mathscr{L}_s g_1)\, g_2 \mathrm{d}\mu_s = -\frac{s}{2Z_s} \int_{\mathbb{R}^d} (\nabla g_1 \cdot \nabla g_2) \mathrm{e}^{-\frac{2f}{s}} \mathrm{d}x = -\frac{s}{2} \int_{\mathbb{R}^d} (\nabla g_1 \cdot \nabla g_2) \mathrm{d}\mu_s.$$

Therefore, $\mathscr{L}_s$ is self-adjoint in $L^2(\mu_s)$ and is non-positive.

C.3 Proof of Lemma 12

For completeness, we show below the original proof of Theorem 12 from Villani (2009) in detail. Let $V_s = \|\nabla f\|^2/s - \Delta f$, then for any $h \in C_c^\infty(\mathbb{R}^d)$ with mean-zero condition

$$\int_{\mathbb{R}^d} h \mathrm{d}\mu_s = 0, \tag{49}$$

we can obtain the following key inequality Deuschel and Stroock (2001)

$$\int_{\mathbb{R}^d} V_s h^2 \mathrm{d}\mu_s \leq s \int_{\mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s. \tag{50}$$

To show (50), note that

$$\begin{aligned}
0 &\leq \int_{\mathbb{R}^d} \left\| \nabla \left(h \mathrm{e}^{-\frac{f}{s}} \right) \right\|^2 \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left\| (\nabla h) \mathrm{e}^{-\frac{f}{s}} - \frac{h}{s} (\nabla f) \mathrm{e}^{-\frac{f}{s}} \right\|^2 \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \|\nabla h\|^2 \mathrm{e}^{-\frac{2f}{s}} \mathrm{d}x + \frac{1}{s} \int_{\mathbb{R}^d} (h^2 \Delta f) \mathrm{e}^{-\frac{2f}{s}} \mathrm{d}x - \left(\frac{1}{s} \right)^2 \int_{\mathbb{R}^d} h^2 \|\nabla f\|^2 \mathrm{e}^{-\frac{2f}{s}} \mathrm{d}x.
\end{aligned}$$

Recognizing $\mu_s \propto \mathrm{e}^{-\frac{2f}{s}}$, this proves (50).

Let $R_{0,s} > 0$ be large enough such that $V_s(x) > 0$ for $\|x\| \geq R_{0,s}$. For $R_s > R_{0,s}$, we can define ϵ_s as

$$\epsilon_s(R_s) := \frac{1}{\inf\{V_s(x) : \ \|x\| \geq R_s\}}. \tag{51}$$

Then $\epsilon(R_s) \to 0$ as $R_s \to \infty$. Furthermore, we assume the R_s is large enough such that

$$\int_{\|x\| \le R_s} \mathrm{d}\mu_s \ge \frac{1}{2}. \tag{52}$$

From the key inequality (50), we obtain that

$$\int_{|x| \ge R_s} h^2 \mathrm{d}\mu_s \le \epsilon(R_s) \left[s \int_{\mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s - (\inf_{x \in \mathbb{R}^d} V_s(x)) \int_{\mathbb{R}^d} h^2 \mathrm{d}\mu_s \right]. \tag{53}$$

Let B_{R_s} be the ball centered at the origin of radius R_s in $\mathbb{R}^d$ and define

$$\mu_{s,R_s} = \left[\int_{|x| \le R_s} \mathrm{d}\mu_s \right]^{-1} \mu_s \mathbf{1}_{|x| \le R_s}.$$

Using the Poincaré inequality in a bounded domain (Evans, 2010, Theorem 1, Chapter 5.8), we get

$$\int_{x \in \mathbb{R}^d} h^2 \mathrm{d}\mu_{s,R_s} \le sC(R_s) \int_{x \in \mathbb{R}^d} \|\nabla h\|^2 \mu_{s,R_s} \mathrm{d}\mu_{s,R_s} + \left(\int_{x \in \mathbb{R}^d} h \mathrm{d}\mu_{s,R_s} \right)^2,$$

where $C(R_s)$ is a constant depending on R_s. Furthermore, using the inequality (52), we obtain

$$\int_{\|x\| \le R_s} h^2 \mathrm{d}\mu_s \le sC(R_s) \int_{\|x\| \le R_s} \|\nabla h\|^2 \mathrm{d}\mu_s + 2 \left(\int_{\|x\| \le R_s} h \mathrm{d}\mu_s \right)^2. \tag{54}$$

Making use of the mean-zero property of h, we have

$$\left(\int_{\|x\| \le R_s} h \mathrm{d}\mu_s \right)^2 = \left(\int_{\|x\| > R_s} h \mathrm{d}\mu_s \right)^2 \le \int_{\|x\| > R_s} h^2 \mathrm{d}\mu_s. \tag{55}$$

Combining (54) and (55), we get

$$\int_{x \in \mathbb{R}^d} h^2 \mathrm{d}\mu_s \le sC(R_s) \int_{x \in \mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s + 3 \int_{\|x\| \ge R_s} h^2 \mathrm{d}\mu_s. \tag{56}$$

Taking (53) and (56) together, we obtain

$$\int_{\mathbb{R}^d} h^2 \mathrm{d}\mu_s \le s[C(R_s) + 3\epsilon(R_s)] \int_{\mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s - 3(\inf_{x \in \mathbb{R}^d} V_s(x)).\epsilon_s(R_s) \int_{x \in \mathbb{R}^d} h^2 \mathrm{d}\mu_s \tag{57}$$

Apparently, from the definition of $\epsilon_s(x)$, we can select $R_s > 0$ large enough such that $1 + 3s(\inf_{x \in \mathbb{R}^d} V_s(x))\epsilon(R_s) > 0$. Then, we can rewrite (57) as

$$\int_{\mathbb{R}^d} h^2 \mathrm{d}\mu_s \le \frac{s}{2} \cdot \frac{2(C(R_s) + 3\epsilon(R_s))}{1 + 3s(\inf_{x \in \mathbb{R}^d} V_s(x))\epsilon(R_s)} \int_{x \in \mathbb{R}^d} \|\nabla h\|^2 \mathrm{d}\mu_s. \tag{58}$$

Finally, using $h - \int_{\mathbb{R}^d} h \mathrm{d}\mu_s$ instead of h in the inequality (58), we prove the desired Poincaré inequality by taking

$$\lambda_s = \frac{1 + 3s(\inf_{x \in \mathbb{R}^d} V_s(x))\epsilon(R_s)}{2(C(R_s) + 3\epsilon(R_s))}.$$

C.4 Proof of Lemma 14

For convenience, we introduce a shorthand:

$$\Pi\left(\frac{g}{s}\right)=\frac{\mathrm{e}^{-\frac{2g}{s}}}{\int_{\mathbb{R}^d}\mathrm{e}^{-\frac{2g}{s}}\,\mathrm{d}x}.$$

Then, we can rewrite the derivative as

$$\begin{aligned}\frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s}&=\frac{\frac{2}{s^2}\int_{\mathbb{R}^d}g^2\mathrm{e}^{-\frac{2g}{s}}\,\mathrm{d}x\int_{\mathbb{R}^d}\mathrm{e}^{-\frac{2g}{s}}\,\mathrm{d}x-\frac{2}{s^2}\left(\int_{\mathbb{R}^d}g\mathrm{e}^{-\frac{2g}{s}}\,\mathrm{d}x\right)^2}{\left(\int_{\mathbb{R}^d}\mathrm{e}^{-\frac{2g}{s}}\,\mathrm{d}x\right)^2}\\&=2\int_{\mathbb{R}^d}\left(\frac{g}{s}\right)^2\Pi\left(\frac{g}{s}\right)dx-2\left(\int_{\mathbb{R}^d}\frac{g}{s}\cdot\Pi\left(\frac{g}{s}\right)dx\right)^2.\end{aligned}$$

Next, we assume that $\zeta_k(x)=x^ke^{-x^\alpha}$, where $\alpha<1$ is a fixed positive constant and $k=1,2$. The facts that $\zeta_k(0)=0$ and $\lim\limits_{x\to+\infty}\zeta_k(x)=0$ give

$$0\le\lim_{s\to0^+}\left(\frac{g}{s}\right)^k\Pi\left(\frac{g}{s}\right)\le\lim_{s\to0^+}\zeta_k\left(\frac{g}{s}\right)=0.$$

Then, by Fatou's lemma, we get

$$\begin{aligned}0\le\liminf_{s\to0^+}\frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s}&\le\limsup_{s\to0^+}\frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s}\\&=2\limsup_{s\to0^+}\int_{\mathbb{R}^d}\left(\frac{g}{s}\right)^2\Pi\left(\frac{g}{s}\right)\mathrm{d}x-2\liminf_{s\to0^+}\left(\int_{\mathbb{R}^d}\frac{g}{s}\cdot\Pi\left(\frac{g}{s}\right)\mathrm{d}x\right)^2\\&\le2\int_{\mathbb{R}^d}\limsup_{s\to0^+}\left(\frac{g}{s}\right)^2\Pi\left(\frac{g}{s}\right)\mathrm{d}x-2\left(\int_{\mathbb{R}^d}\liminf_{s\to0^+}\frac{g}{s}\cdot\Pi\left(\frac{g}{s}\right)\mathrm{d}x\right)^2=0.\end{aligned}$$

The proof is complete.