Stochastic Gradient Descent in Nonconvex Optimization: Continuous-Time Dynamics and the Role of Learning Rates

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 9, 2024

Abstract

Stochastic Gradient Descent (SGD) has become a cornerstone of machine learning optimization, especially in nonconvex settings. Despite its empirical success, the theoretical understanding of the influence of learning rates on convergence remains incomplete. This paper provides a novel theoretical framework by leveraging stochastic differential equations (SDEs) and Schrödinger operators to study the continuous-time behavior of SGD. We demonstrate how the learning rate impacts convergence dynamics, with a particular focus on nonconvex optimization problems. Our results provide new insights into the scheduling of learning rate decay and its effectiveness in escaping saddle points. We also establish linear convergence rates to stationarity in specific settings.

Keywords: Stochastic Gradient Descent, Learning Rate, Nonconvex Optimization, Schrödinger Operators, Convergence Dynamics.

1 Introduction

Gradient-based optimization algorithms, particularly Stochastic Gradient Descent (SGD), have been pivotal in driving recent advancements in machine learning. Their simplicity, efficiency, and scalability make them essential tools for solving large-scale problems, especially in the training of deep neural networks. While classical optimization theory has focused primarily on convex problems, many real-world applications in machine learning require solving nonconvex optimization tasks, where saddle points and local minima pose significant challenges for optimization algorithms.

SGD has been particularly successful in nonconvex settings due to its favorable runtime properties and its ability to generalize well across a variety of problems. However, despite its empirical success, the theoretical understanding of SGD, especially regarding how certain hyperparameters influence its performance, remains incomplete. Chief among these is the learning rate, which controls the size of each step taken during optimization. The learning rate profoundly impacts both convergence speed and the ability to escape poor local optima. Practically, learning rate schedules—such as decay schedules or cyclical patterns—are often employed to improve performance, yet the precise mathematical relationship between the learning rate and SGD's behavior in nonconvex settings remains elusive.

In this paper, we seek to bridge this gap by providing a rigorous theoretical analysis of the impact of learning rates on the convergence of SGD in nonconvex optimization. By leveraging stochastic differential equations (SDEs) as continuous-time surrogates for SGD, we are able to explore the deeper dynamics of the optimization process. Additionally, we utilize tools from the theory of Schrödinger operators to analyze the convergence properties of SGD in nonconvex landscapes, offering new insights into the role of the learning rate in these challenging settings.

Overview of Contributions In this paper, we make the following contributions:

- We establish a theoretical framework for analyzing the effect of learning rates in nonconvex optimization using stochastic differential equations (SDEs).
- We demonstrate that SGD exhibits linear convergence to stationarity in certain settings and explain how the learning rate influences this rate.
- We use the framework of Schrödinger operators to provide insight into the role of learning rates in escaping saddle points, a critical challenge in nonconvex optimization.
- Our results provide new guidelines for setting learning rate decay schedules, with theoretical backing for strategies used in deep learning.

Mathematically, 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 of this 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 compute 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

$$x_{k+1} = x_k - s\widetilde{\nabla} f(x_k) = x_k - s\nabla f(x_k) - s\xi_k \tag{1.1}$$

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 [Bot10].

The learning rate plays an essential role in determining the performance of SGD and many of the practical variants of SGD [Ben12].[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 [Zei12, KB14]. 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 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,

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

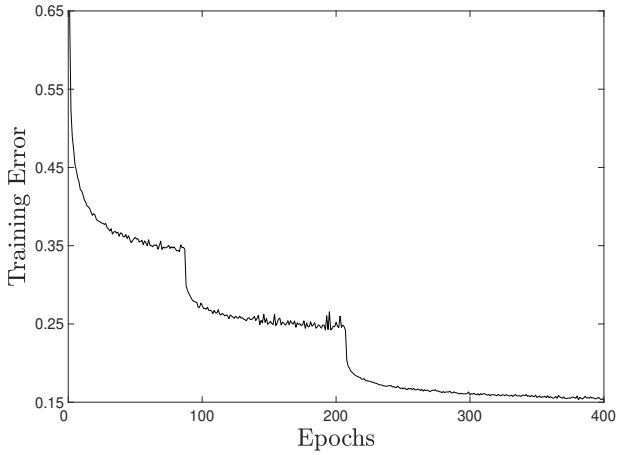

Figure 1. Training error using SGD with mini-batch size 32 to train an 8-layer convolutional neural network on CIFAR-10 [Kri09]. The first 90 epochs use a learning rate of $s = 0.006$, the next 120 epochs use $s = 0.003$, and the final 190 epochs use $s = 0.0005$. Note that the training error decreases as the learning rate s decreases and a smaller s leads to a larger number of epochs for SGD to reach a plateau. See [HZRS16] for further investigation of this phenomenon.

a technique that is used in training deep neural networks (see, e.g., [HZRS16, BCN18, SS19]). 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 [YLWJ19, LWM19]. Accordingly, strategies for setting learning rate decay schedules are generally adhoc 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 [SBC16, Jor18, SDJS18]. Specifically, in the case of SGD, we study stochastic differential equations (SDEs) as surrogates for discrete stochastic optimization methods (see, e.g., [KY03, LTE17, KB17, COO$^+$18, DJ19]). 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 and, for the time being, 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 (1.1), we get

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

[2]Although a Brownian motion is not differentiable, the formal notation $\mathrm{d}W(t)/\mathrm{d}t$ can be given a rigorous interpretation [Eva12, Vil06].

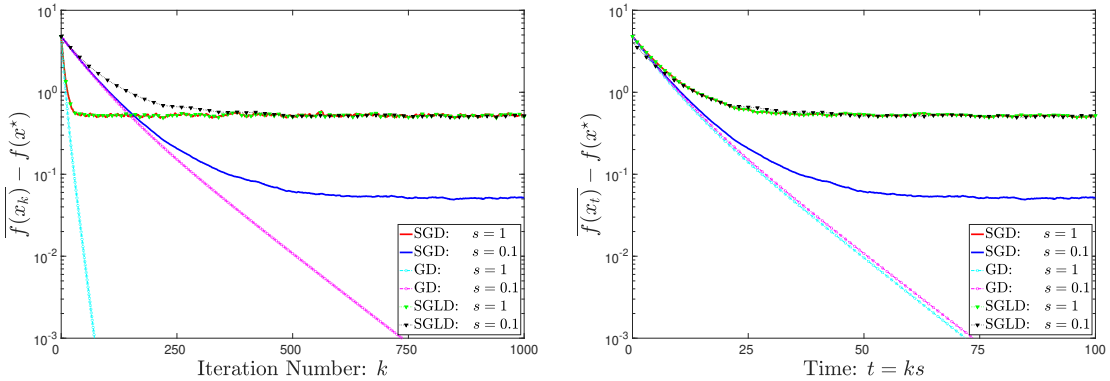

Figure 2. Illustrative examples showing distinct behaviors of GD, SGD, and SGLD. The y-axis displays the optimization error $\overline{f(x_k)} - f(x^\star)$, where $f(x^\star)$ denotes the minimum value of the objective and in the case of SGD and SGLD $\overline{f(x_k)}$ denotes an average over 1000 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.

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:

$$\mathrm{d}X_s = -\nabla f(X_s)\mathrm{d}t + \sqrt{s}\mathrm{d}W \tag{1.2}$$

where the initial condition is the same value x_0 as its discrete counterpart. This SDE has been shown to be a valid approximating surrogate for SGD in earlier work [KY03, CS18]. 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 [SKYL17]). Throughout this paper we focus on (1.2) 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 (1.2), 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 in gradient descent (GD) and stochastic gradient Langevin dynamics (SGLD) when one considers their limiting 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:

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

[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 (1.2), 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.

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 (1.2). 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.

(i) **Linear convergence to stationarity.** 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, we prove that the solution $X_s(t)$ to the lr-dependent SDE obeys

$$\mathbb{E}\, f(X_s(t)) - f^\star \le \epsilon(s) + C(s)\mathrm{e}^{-\lambda_s t} \tag{1.3}$$

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 can show that $\epsilon(s)$ decreases monotonically to zero as $s \to 0$. This bound can be carried over to the discrete case by a uniform approximation between SGD and the lr-dependent SDE (1.2). Specifically, the term $C(s)\mathrm{e}^{-\lambda_s t}$ becomes $C(s)\mathrm{e}^{-\lambda_s ks}$, 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, note that while GD can be trapped in a saddle point or a local minimum, SGD can efficiently escape saddle points, 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 [JGN$^+$17, LSJR16].

(ii) **Distinctions between convexity and nonconvexity.** 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

[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.

[5] Roughly speaking, stationarity refers to the distribution of $X_s(t)$ in the limit $t \to \infty$. See a more precise definition in Section 3.

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]

$$\lambda_s \asymp \mathrm{e}^{-\frac{2H_f}{s}} \tag{1.4}$$

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.

For comparison, λ_s is equal to μ if f is μ-strongly convex for $\mu > 0$, regardless of the learning rate s. 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 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 (1.3). 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 (1.2) are as follows. We establish the linear convergence via a Poincaré-type inequality that is due to Villani [Vil09]. 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. 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 [SKYL17, KMN$^+$16], the authors uncovered various tradeoffs linking the learning rate and the mini-batch size. Moreover, [JKA$^+$17, JKB$^+$18] 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 [LWM19, YLWJ19]. From a topological perspective, [DDC19] 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 [LH16, Smi17] (see also the review [Sun19]), with recent theoretical justification [LA19].

[6] We write $a_m \asymp b_m$ if there exist positive constants c and c' such that $cb_m \le a_m \le c'b_m$ for all m.

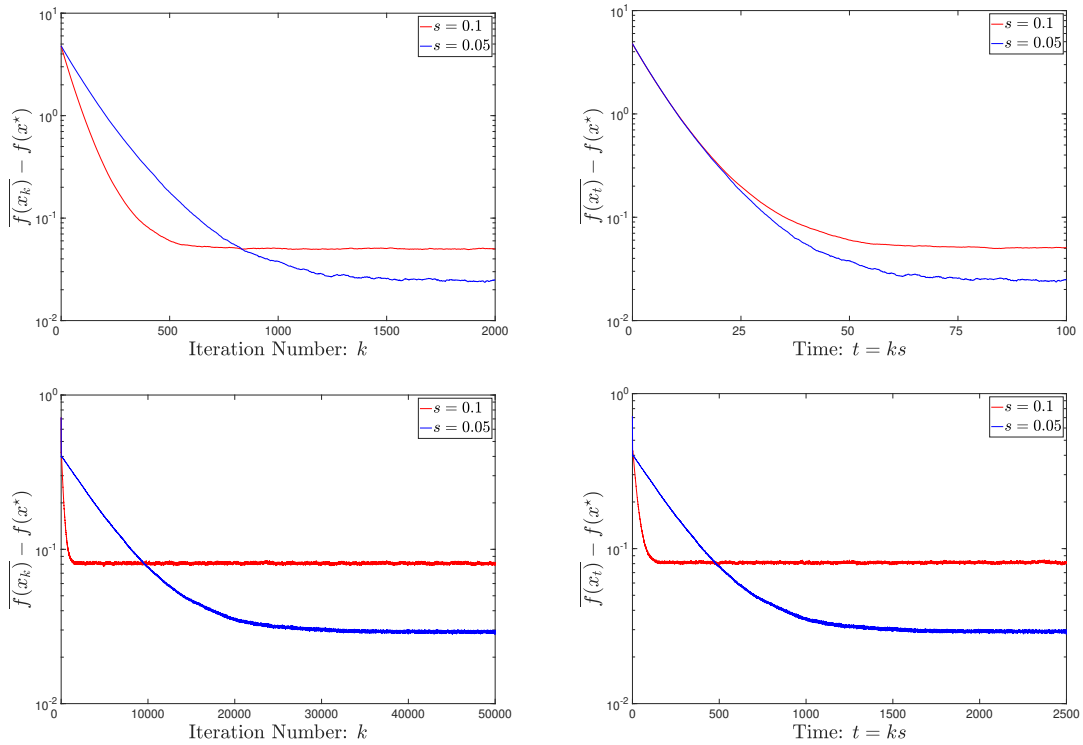

Figure 3. The dependence of the optimization dynamics of SGD on the learning rate *differs* between convex objectives and nonconvex objectives. The learning rate is set to either $s = 0.1$ or $s = 0.05$. The two top plots consider minimizing a convex function $f(x_1, x_2) = 5 \times 10^{-2} x_1^2 + 2.5 \times 10^{-2} x_2^2$, 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.

In a different direction, there has been a flurry of activity in using dynamical systems to analyze discrete optimization methods. For example, [SBC16, WWJ16, SDJS18] derived ODEs for modeling Nesterov's accelerated gradient methods and used the ODEs to understand the acceleration phenomenon (see the review [Jor18]). In the stochastic setting, this approach has been recently pursued by various authors [COO$^+$18, CS18, MHB16, LSJR16, CH19, LTE17] 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, [HRB08]).

Our work is motivated in part by the recent progress on Langevin dynamics, in particular in nonconvex settings [Vil09, Pav14, HKN04, BGK05]. 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 [Hwa80, RRT17, ZLC17]. Compared to these results, however, the present paper is distinct in that our analysis provides a more concise and sharp delineation of the convergence rate based on geometric properties of the objective function.

1.3 Organization

The paper is organized as follows: In Section 2, we introduce the basic assumptions and techniques employed throughout the paper. Section 3 presents our main theoretical results, including the convergence rates of SGD and their dependence on the learning rate. In Section 4, we explore the practical implications of using a large initial learning rate followed by a sequence of decreasing learning rates for training neural networks. Section 5 provides formal proofs of the linear convergence and Section 6 further specifies the rate of convergence by analyzing the spectrum of the Schrödinger operator. Technical details are deferred to the appendices. Finally, we conclude the paper in Section 7 with potential 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 2.1 (Confining condition [Pav14, MV99])**.** 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 and, indeed, 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 (1.2) 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 3.5. Recall that the lr-dependent SDE (1.2) 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 Fokker–Planck–Smoluchowski equation

$$\frac{\partial \rho_s}{\partial t} = \nabla \cdot (\rho_s \nabla f) + \frac{s}{2}\Delta \rho_s \tag{2.1}$$

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 Fokker–Planck–Smoluchowski equation from the lr-dependent SDE (1.2) 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

$$\mu_s = \frac{1}{Z_s}\mathrm{e}^{-\frac{2f}{s}} \tag{2.2}$$

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 2.2 (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 (1.2) 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 Fokker–Planck–Smoluchowski equation (2.1).

The proof of Lemma 2.2 is shown in Appendix A.4. For more information, Lemma 5.2 in Section 5 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 2.3 (Villani condition [Vil09])**.** 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 the SGD iterates in the training process are bounded and this condition is essentially concerned with the function at infinity.

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

$$\mathbb{E}\, f(X_s(t)) - f^\star \le \epsilon(s) + D(s)\mathrm{e}^{-\lambda_s t} \tag{3.1}$$

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 5.2). 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 3.1. 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 $t \ge 0$, 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 3.1 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 write henceforth $\epsilon(s) := \mathbb{E}\, f(X_s(\infty)) - f^\star$.

Proposition 3.2. 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

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

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

The two propositions are proved in Section 5. The proof of Theorem 1 is a direct consequence of Proposition 3.1 and Proposition 3.2. More precisely, the two propositions taken together give

$$\mathbb{E}\, f(X_s(t)) - f^\star \le O(s) + C(s)\mathrm{e}^{-\lambda_s t} \tag{3.2}$$

for a bounded learning rate s.

Taken together, these results offer insights into the phenomena observed in Figure 1. In particular, Proposition 3.1 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 3.2 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 Figures 1, 2, and 3. For comparison, this property is not observed in GD and SGLD.

The following result gives the iteration complexity of SGD in its continuous-time formulation.

Corollary 3.3. Under the assumptions of Proposition 3.2, for any $\epsilon > 0$, if the learning rate $s \le \min\{\epsilon/(2A), S\}$ and $t \ge \frac{1}{\lambda_s} \log \frac{2C(s)\|\rho - \mu_s\|_{\mu_s^{-1}}}{\epsilon}$, then

$$\mathbb{E}\, f(X_s(t)) - f^\star \le \epsilon$$

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

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

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 [BGL13].[7]

Proposition 3.4. In addition to the assumptions of Theorem 1, assume that the objective f is a μ-strongly convex function. Then, λ_s in (3.1) 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, $\nabla f(x) = 0$).[8]

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.[9] Then the constant λ_s in (3.1) satisfies

$$\lambda_s = (\alpha + o(s))\mathrm{e}^{-\frac{2H_f}{s}} \tag{3.3}$$

for $0 < s \le s_0$, 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 (3.1) 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 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).

[7] In fact, we can obtain a tighter log-Sobolev inequality for convergence of the probability densities in $L^1(\mathbb{R}^d)$, as is shown in Appendix B.2.

[8] See Section 6.2 for a discussion of Morse functions. Note that (infinitely differentiable) strongly convex functions are Morse functions.

[9] 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.

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 $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ 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 3.5. 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 (1.2) 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

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

for any fixed $T > 0$.

We note that there exists a sharp bound on $B(T)$ in [BT96]. For completeness, we also remark that the convergence can be strengthened to the strong sense:

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

This result has appeared in [Mil75, Tal82, PT85, Tal84, KP92] and we provide a self-contained proof in Appendix B.3.

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, the following two conclusions hold:

(a) For any $T > 0$, the iterates of SGD with learning rate $0 < s \leq 1/L$ satisfy

$$\mathbb{E}\, f(x_k) - f^\star \leq (A + B(T))s + C\, \|\rho - \mu_s\|_{\mu_s^{-1}}\, \mathrm{e}^{-s\lambda_s k} \tag{3.4}$$

for all $k \leq T/s$, where λ_s is the exponential decay constant in (3.1), A as in Proposition 3.2 depends only on $1/L$ and f, $C = C_{1/L}$ is as in Proposition 3.1, and $B(T)$ depends only on the time horizon T and the Lipschitz constant L.

(b) If f is a Morse function with at least two local minima, with λ_s appearing in (3.4) being given by (3.3), and if f is μ-strongly convex then $\lambda_s = \mu$.

Theorem 3 follows as a direct consequence of Theorem 1 and Proposition 3.5. Note that the second part of Theorem 3 is simply a restatement of Theorem 2 and Proposition 3.4. 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 (3.4) is always an increasing function of s even for the strongly convex case, where λ_s itself is constant.

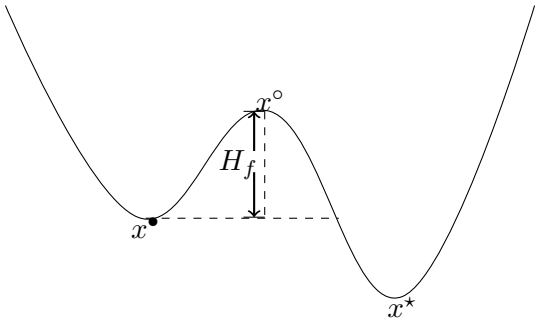

Figure 4. A one-dimensional nonconvex function f. The height difference between x° and $x^\bullet$ in this special case is the Morse saddle barrier H_f. See the formal definition in Definition 6.5.

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°.[10] We use this toy example to gain insight into the expression (3.3) for the exponential decay constant λ_s; deferring the rigorous derivation of this number in the general case to Section 6.

From (3.1) it suggests that the lr-dependent SDE (1.2) 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 $x \le x^\circ - \nu$ for a very small positive ν and $\theta > 0$. Accordingly, the lr-dependent SDE (1.2) 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

$$\mathbb{E}\,\tau_{x^\circ} \approx \frac{\sqrt{\pi s}}{(x^\circ - x^\bullet)\theta\sqrt{\theta}}\,\mathrm{e}^{\frac{2}{s}\cdot\frac{1}{2}\theta(x^\circ - x^\bullet)^2} \approx \frac{\sqrt{\pi s}}{(x^\circ - x^\bullet)\theta\sqrt{\theta}}\,\mathrm{e}^{\frac{2H_f}{s}} \tag{3.5}$$

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 (3.5), 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 (3.3).

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

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

[10] 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.

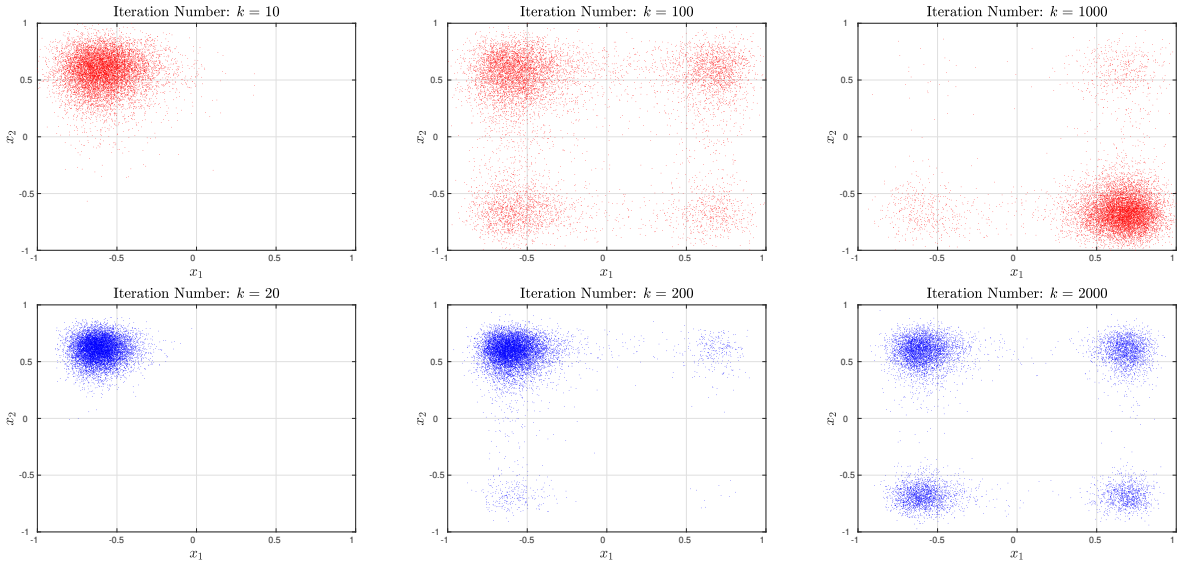

Figure 5. Scatter plots of the iterates $x_k \in \mathbb{R}^2$ of SGD for minimizing the nonconvex function in Figure 3. This function has four local minima, of which the bottom right one is the gloabl minimum. Each column corresponds to the same value of $t = ks$, and the first row and second row correspond 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 $(-0.9, 0.9)$.

See, for example, [FW12, Pav14]. Furthermore, we demonstrate the view from the theory of *viscosity solution* and *singular perturbation* in Appendix B.4.

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 toward understanding its benefits [YLWJ19, LWM19]. 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 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,

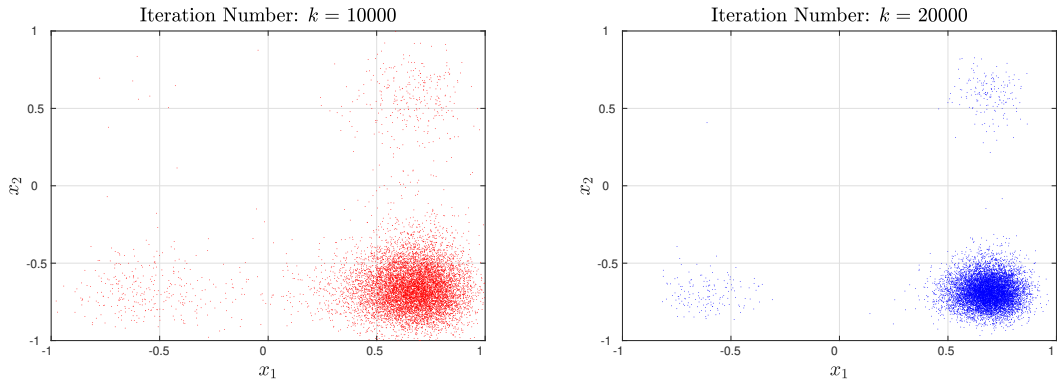

Figure 6. The same setting as in Figure 5. Both plots correspond to the same value of $t = ks = 1000$.

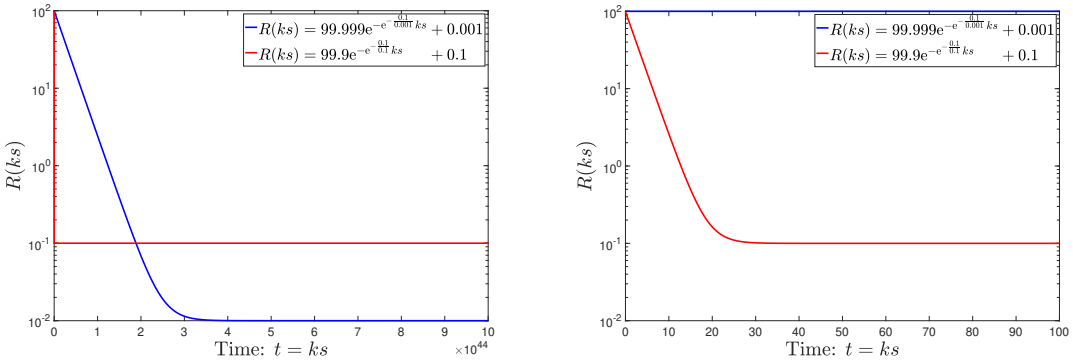

Figure 7. Idealized risk function of the form $R(t) = as + be^{-e^{\frac{c}{s}t}}$ with the identification $t = ks$, which is adapted from (3.2). The parameters are set as follows: $a = 1, b = 100 - s, c = 0.1$, and the learning rate is $s = 0.1$ or 0.001. The right plot is a locally enlarged image of the left.

however, enable a concrete and crisp understanding of the vital importance of nonconvexity in this setting. Motivated by (3.2), we consider an idealized risk function of the form $R(t) = as + be^{-\lambda_s t}$, with λ_s set to $e^{-c/s}$, where a, b, and c are positive constants for simplicity as opposed to the non-constants in the upper bound in (3.1). 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 $k_{0.001} \approx 2.5 \times 10^{47}$. 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

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

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, $\epsilon(s) \equiv$

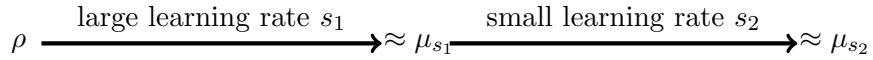

Figure 8. Learning rate decay. The first phase uses a larger learning rate s_1, at the end of which the SGD iterates are approximately distributed as μ_{s_1}. The second phase uses a smaller learning rate s_2 and at the end the distribution of the SGD iterates roughly follows μ_{s_2}.

$\mathbb{E}\, f(X_s(\infty)) - f^\star$, which is indeed the claim of Proposition 3.2. Leveraging Proposition 3.1, 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

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

for a small $\delta > 0$. In words, the lr-dependent SDE (1.2) 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 3.1 it follows that (recall that ρ is the initial distribution):

$$T_s^\delta \le \frac{1}{\lambda_s} \log \frac{C(s)\, \|\rho - \mu_s\|_{\mu_s^{-1}}}{\delta\epsilon(s)} = \frac{\mathrm{e}^{\frac{2H_f}{s}}}{\gamma + o(s)} \log \frac{C(s)\, \|\rho - \mu_s\|_{\mu_s^{-1}}}{\delta\epsilon(s)} \tag{4.1}$$

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 $\|\rho - \mu_s\|_{\mu_s^{-1}}$ in (4.1) for the second phase of learning rate decay is approximately

$$\|\mu_{s_1} - \mu_{s_2}\|_{\mu_{s_2}^{-1}} = \left(\int (\mu_{s_1} - \mu_{s_2})^2 \mu_{s_2}^{-1} \mathrm{d}x\right)^{\frac{1}{2}} = \left(\int \frac{\mu_{s_1}^2}{\mu_{s_2}} \mathrm{d}x - 1\right)^{\frac{1}{2}} \tag{4.2}$$

Both μ_{s_1} and μ_{s_2} are decreasing functions of f and, therefore, have the same modes. As a consequence, the integral of $\mu_{s_1}^2/\mu_{s_2}$ is small by appeal 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, $\|\rho - \mu_{s_2}\|_{\mu_{s_2}^{-1}}$ 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.

5 Proof of the Linear Convergence

In this section, we prove Proposition 3.1 and Proposition 3.2, leading to a complete proof of Theorem 1.

5.1 Proof of Proposition 3.1

To better appreciate the linear convergence of the lr-dependent SDE (1.2), as established in Proposition 3.1, 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 3.1.

Convergence without a rate. Recall that we use ρ to denote the initial probability density in the space $L^2(\mu_s^{-1})$. Superficially, it seems that the most natural space for probability densities is $L^1(\mathbb{R}^d)$. However, we prefer to work in $L^2(\mu_s^{-1})$ since this function space has certain appealing properties that allow us to obtain the proof of the desired convergence results for the lr-dependent SDE. 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 5.1. 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 5.2. 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 (2.1) converges in $L^2(\mu_s^{-1})$ to the Gibbs invariant distribution μ_s, which is specified by (2.2).

Note that the existence and uniqueness of $\rho_s(t,\cdot)$ is ensured by Lemma 2.2. The convergence guarantee on $\rho_s(t,\cdot)$ in Lemma 5.2 relies heavily on the following lemma (Lemma 5.3). 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

$$\frac{\partial h_s}{\partial t} = -\nabla f \cdot \nabla h_s + \frac{s}{2}\Delta h_s \tag{5.1}$$

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

$$\mathscr{L}_s = -\nabla f \cdot \nabla + \frac{s}{2}\Delta \tag{5.2}$$

has a crucial property, as stated in the following lemma. Its proof is postponed to Appendix C.2.

Lemma 5.3. The linear operator $\mathscr{L}_s$ in (5.2) is self-adjoint and nonpositive in $L^2(\mu_s)$. Explicitly, for any g_1, g_2, this operator obeys

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

Proof of Lemma 5.2. We have

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

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

where the last equality is due to (5.1). Next, we proceed by making use of Lemma 5.3:

$$2\int_{\mathbb{R}^d}(h_s-1)\mathscr{L}_s(h_s-1)\mathrm{d}\mu_s = -s\int_{\mathbb{R}^d}\nabla(h_s-1)\cdot\nabla(h_s-1)\mathrm{d}\mu_s$$
$$= -s\int_{\mathbb{R}^d}\|\nabla h_s\|^2\mathrm{d}\mu_s \leq 0 \tag{5.3}$$

Thus, $\|\rho_s(t,\cdot)-\mu_s\|^2_{\mu_s^{-1}}$ 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 $L^2(\mu_s^{-1})$. □

Linear convergence. We turn towards 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 3.1.

Lemma 5.4 (Theorem A.1 in [Vil09])**.** 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 above 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 5.4. In addition, while the constant in the Poincaré inequality in general depends on the dimension (see, for example, [Eva10, Theorem 1, Chapter 5.8]), λ_s in Lemma 5.4 is completely determined by geometric properties of the objective f. See details in Section 6.

Importantly, Lemma 5.4 allows us to obtain the following lemma, from which the proof of Proposition 3.1 follows readily. The proof of this lemma is given at the end of this subsection.

Lemma 5.5. Under the assumptions of Proposition 3.1, $\rho_s(t,\cdot)$ converges to the Gibbs invariant distribution μ_s in $L^2(\mu_s^{-1})$ at the rate

$$\|\rho_s(t,\cdot)-\mu_s\|_{\mu_s^{-1}}\leq \mathrm{e}^{-\lambda_s t}\|\rho-\mu_s\|_{\mu_s^{-1}} \tag{5.4}$$

Proof of Proposition 3.1. Using Lemma 5.5, we get

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

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 5.5.

Proof of Lemma 5.5. It follows from (5.3) 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 5.4 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

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

Integrating both sides yields (5.4), as desired.

□

5.2 Proof of Proposition 3.2

Next, we turn to the proof of Proposition 3.2. We first state a technical lemma, deferring its proof to Appendix C.4.

Lemma 5.6. Under the assumptions of Proposition 3.2, the excess risk at stationarity $\epsilon(s)$ satisfies

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

Using Lemma 5.6, we now finish the proof of Proposition 3.2.

Proof of Proposition 3.2. 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

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

As a consequence, $\epsilon(0) = 0$. Lemma 5.6 shows that for any $S > 0$, there exists $A = A_S$ such that $0 \leq \frac{\mathrm{d}\epsilon(s)}{\mathrm{d}s} \leq A$ for all $0 \leq s \leq S$. This fact, combined with $\epsilon(0) = 0$, immediately gives $\epsilon(s) \leq As$ for all $0 \leq s \leq S$.

□

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 (1.2) and a Schrödinger operator. 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 (2.1) as

$$\frac{\partial \psi_s}{\partial t} = \frac{s}{2}\Delta\psi_s - \left(\frac{\|\nabla f\|^2}{2s} - \frac{\Delta f}{2}\right)\psi_s = -\frac{-s\Delta + V_s}{2}\psi_s \tag{6.1}$$

with the initial condition $\psi_s(0, \cdot) = \frac{\rho}{\sqrt{\mu_s}} \in L^2(\mathbb{R}^d)$. This is a Schrödinger equation with the associated operator $-s\Delta + V_s$, 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 (2.2), it is not hard to show that $\sqrt{\mu_s}$ is the unique eigenfunction of $-s\Delta + V_s$ with a corresponding eigenvalue of zero. Using this fact, from the proof of Lemma 5.5, we get

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

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 $L^2(\mathbb{R}^d)$ [HS12].

Lemma 6.1 (Theorem 10.7 in [HS12])**.** 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 6.1 allow us to order the eigenvalues of $-s\Delta + V_s$ in $L^2(\mathbb{R}^d)$ as

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

A crucial fact from this representation is that the exponential decay constant λ_s in Theorem 5.5 can be set to

$$\lambda_s = \frac{1}{2}\zeta_{s,1} \tag{6.2}$$

To see this, note that $\psi_s(t,\cdot) - \sqrt{\mu_s}$ also satisfies (6.1) 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.[12] 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 5.5.

As a consequence of this discussion, we seek to study the Fokker–Planck–Smoluchowski equation (2.1) by analyzing the spectrum of the linear Schrödinger operator (6.1), especially its smallest

[12] 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,

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

positive eigenvalue $\delta_{s,1}$. To facilitate the analysis, a crucial observation is that this Schrödinger operator is equivalent to the *Witten-Laplacian*,

$$\Delta_f^s := s(-s\Delta + V_s) = -s^2\Delta + \|\nabla f\|^2 - s\Delta f \tag{6.3}$$

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 [HN05, BGK05, Nie04, AK99], 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 [HHS11, Mic19].

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.[13] 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 any index-1 saddle point x, it is not hard to show 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. Using this fact, we give the following definition.

Definition 6.2. 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

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

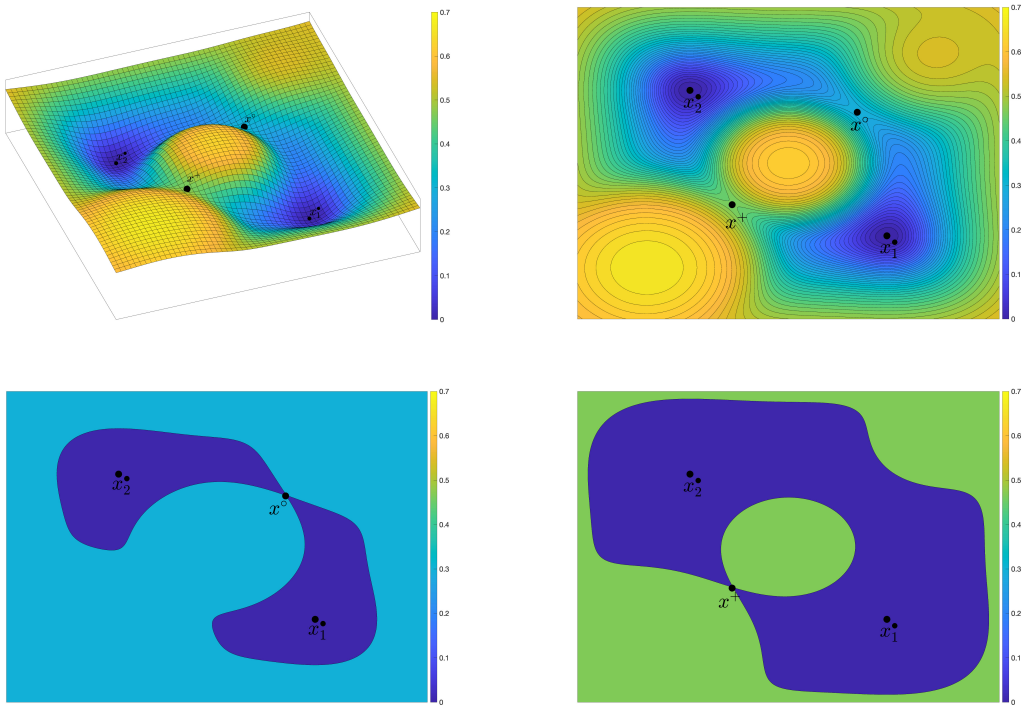

Figure 9. The landscape of a two-dimensional nonconvex Morse function. Here, $x_1^\bullet$ and $x_2^\bullet$ denote two local minima. Both x° and x^+ are index-1 saddle points, but only the former is an index-1 separating saddle point since $f(x^\circ) < f(x^\bullet)$. In the two bottom plots, the deep 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.

index-1 separating saddle point and writes $\mathcal{X}^\circ$ for the set of all these points. To give a geometric interpretation of Definition 6.2, 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.

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 $\mathcal{V} = \{\nu_1, \dots, \nu_I\}$ and sort these values as

$$+\infty = \nu_0 > \nu_1 > \cdots > \nu_I \tag{6.4}$$

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

Next, we follow [HHS11] and define a type of connected components of sublevel set.

Definition 6.3. 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 6.2 for more details.

With the preparatory notions above in place, we describe the following procedure for labeling index-1 separating saddle points and local minima [HHS11]. See Figure 10 for an illustration of this process.

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 1 (Generic case [HHS11])**.** *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.*

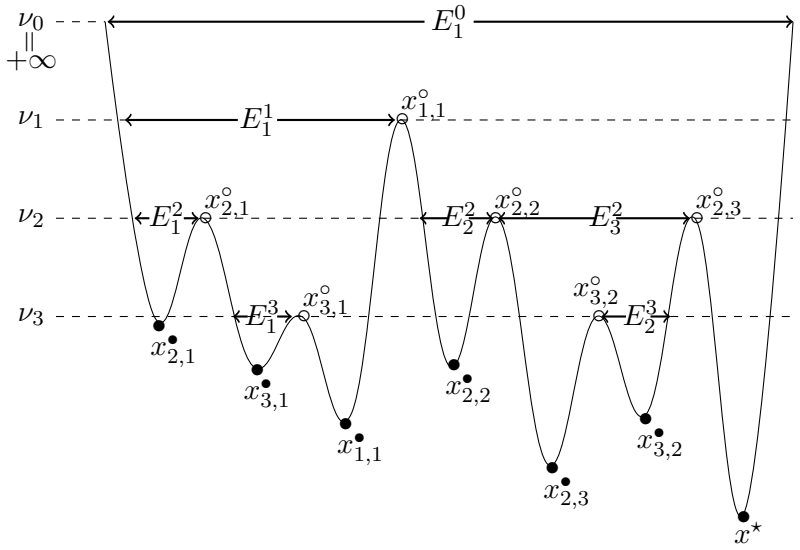

Figure 10. A generic one-dimensional Morse function. The labeling process gives rise to a one-to-one correspondence between the local minimum $x_{ij}^\bullet$ and the index-1 separating saddle point $x_{i,j}^\circ$ (which are also local maxima) for all i, j.

- *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 1, 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

$$f(x_0^\circ) - f(x_0^\bullet) > f(x_1^\circ) - f(x_1^\bullet) \geq \ldots \geq f(x_{n^\bullet-1}^\circ) - f(x_{n^\bullet-1}^\bullet) \tag{6.5}$$

where $f(x_0^\circ) - f(x_0^\bullet) = f(\infty) - f(x^\star) = +\infty$. A detailed description of this bijection is given in [HHS11, 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 (6.3). Recall that the nonconvex Morse function f satisfies the confining condition and the Villani condition.

Proposition 6.4 (Theorem 1.2 in [HHS11])**.** Under Assumption 1 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

$$\gamma_\ell = \frac{|\eta_d(x_\ell^\circ)|}{\pi} \left(\frac{\det(\nabla^2 f(x_\ell^\bullet))}{-\det(\nabla^2 f(x_\ell^\circ))} \right)^{\frac{1}{2}} \tag{6.6}$$

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

Using Proposition 6.4 in conjunction with the simple relationship between the exponential decay constant and the spectrum of the Schrödinger operator/Witten-Laplacian (6.2), 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 6.5. 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 of Theorem 2 in the generic case. By Proposition 6.4, 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 (3.3), 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 1. This simplification is due to [Nie04].

Assumption 2 (Simplified generic case [Nie04])**.** *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 6.4.

Corollary 6.6 (Theorem 3.1 in [Nie04])**.** Under Assumption 2 and the assumptions of Theorem 2, Proposition 6.4 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 1. To violate the generic assumption, for example, we can change the objective value $f(x_{3,1}^\bullet)$ to $f(x_{1,1}^\bullet)$ or change $f(x_{3,2}^\bullet)$ to $f(x_{2,3}^\bullet)$ in Figure 10. In this situation, the first condition in Assumption 1 is not satisfied. Alternatively, if the objective value at $x_{3,1}^\circ$ is changed to $f(x_{2,1}^\circ)$, the second condition in Assumption 1 is not met. Figure 11 presents an example of a degenerate Morse function.

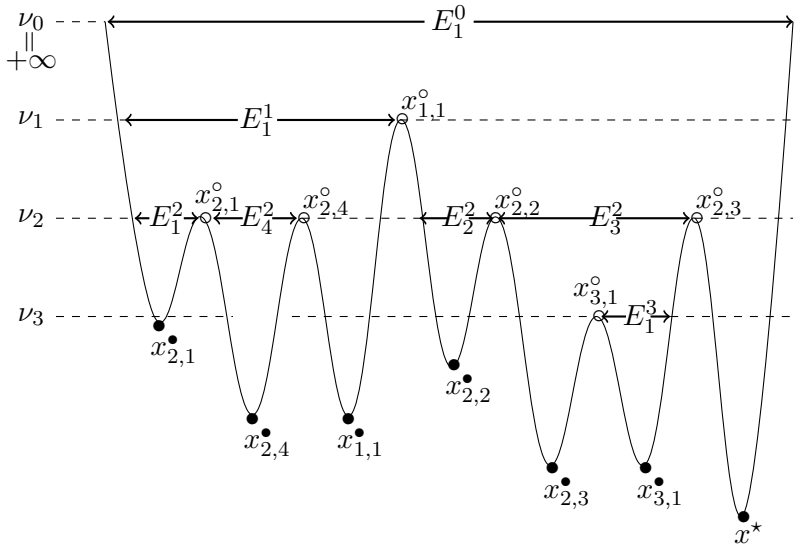

Figure 11. A degenerate one-dimensional Morse function. The labeling of its index-1 separating saddle points $x_{i,j}^\circ$ and local minima $x_{i,j}^\bullet$ is not unique. Nevertheless, the labeling process gives a unique one-to-one correspondence between the function values at the two types of points.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

The main challenge in the degenerate case is the lack of uniqueness of the pairs $(x_\ell^\circ, x_\ell^\bullet)$ 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 $(f(x_\ell^\circ), f(x_\ell^\bullet))$ obeying

$$f(\infty) - f(x^\star) = f(x_0^\circ) - f(x_0^\bullet) > f(x_1^\circ) - f(x_1^\bullet) \geq \ldots \geq f(x_{n^\bullet-1}^\circ) - f(x_{n^\bullet-1}^\bullet)$$

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 6.4 to the degenerate case, which is adapted from Theorem 2.8 of [Mic19].

Proposition 6.7 (Theorem 2.8 in [Mic19])**.** Assume that the assumptions of Theorem 2 are satisfied but not Assumption 1. 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

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

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 6.4 and Proposition 6.7 give a full proof of Theorem 2. As is clear, the Morse saddle barrier in Definition 6.5 for the degenerate case is set to $H_f = H_{f,1}$. For completeness, we remark that this result applies to Assumption 1, 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 (6.6). As such, Proposition 6.4 is implied by Proposition 6.7.

7 Discussion

In this paper, we have provided a theoretical analysis of the impact of the learning rate on Stochastic Gradient Descent (SGD) in nonconvex optimization. By introducing a learning-rate-dependent stochastic differential equation (LR-dependent SDE), we utilized modern tools from diffusion theory, particularly the spectral analysis of diffusion operators, to examine the continuous-time behavior of SGD. Our analysis demonstrated that the solution to the SDE converges linearly to stationarity under certain regularity conditions, and we derived an explicit expression for the linear convergence rate with clear dependence on the learning rate for nonconvex Morse functions.

We uncovered a critical distinction between convex and nonconvex optimization problems. For strongly convex functions, the linear convergence rate remains constant, while in nonconvex settings, it decreases rapidly as the learning rate approaches zero. This highlights the pivotal role that noise in the gradients plays in nonconvex optimization, contrasting with its more limited role in convex problems. Additionally, our results provide a theoretical justification for the common practice of using a large initial learning rate and gradually decaying it during the training of neural networks, offering insights into why learning rate decay is particularly effective for nonconvex objectives.

Future Directions We propose several directions for future research to consolidate and extend the framework for analyzing stochastic optimization methods via SDEs. A pressing question is to better characterize the gap between the stationary distribution of the lr-dependent SDE and that of the discrete SGD [Kro93, Pav14, DDB17]. Explicitly, can we improve the upper bound in Proposition 3.5? A related question is whether Theorem 3 can be improved to $\mathbb{E}\, f(x_k) - f^\star \le O(s + (1-\lambda_s s)^k)$, with the hidden coefficients having less dependence on the time horizon ks. A possible approach to overcoming this difficulty in the discrete regime is to obtain a discrete version of the Poincaré inequality in $\mathbb{R}^d$ (Lemma 5.4). From a different angle, it is noteworthy that $(s/2)\Delta\rho_s$ in the Fokker–Planck–Smoluchowski equation (2.1) corresponds to vanishing viscosity in fluid mechanics. Appendix B.4 presents several open problems from this viewpoint. To widen the scope of this framework, it is important to extend our results to the setting where the gradient noise is heavy-tailed [SSG19].

From a practical standpoint, our work offers several promising avenues for future research in deep learning. First, a seemingly straightforward direction is 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 [BCN18, Smi17]. More broadly, it is of great interest to use SDEs to study and improve on practical variants of SGD, including RMSProp and Adam [TH12, KB14]. Second, our results would likely to 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, 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? Finally, we wonder if the lr-dependent SDE might give insights into generalization properties of neural networks such as local elasticity [HS20] and implicit regularization [ZBH$^+$16, GLSS18].

References

[AK99] V. I. Arnol'd and B. A. Khesin. *Topological Methods in Hydrodynamics*, volume 125. Springer Science & Business Media, 1999.

[Arn12] V. Arnol'd. *Geometrical Methods in the Theory of Ordinary Differential Equations*. Springer Science & Business Media, 2012.

[Arn13] V. Arnol'd. *Mathematical Methods of Classical Mechanics*. Springer Science & Business Media, 2013.

[BCN18] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[Ben12] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer, 2012.

[BGK05] 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.

[BGL13] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer Science & Business Media, 2013.

[Bot10] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[BT96] 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.

[CEL84] 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.

[CF99] G.-Q. Chen and H. Frid. Vanishing viscosity limit for initial-boundary value problems for conservation laws. *Contemporary Mathematics*, 238:35–51, 1999.

[CH19] K. Caluya and A. Halder. Gradient flow algorithms for density propagation in stochastic systems. *IEEE Transactions on Automatic Control*, 2019.

[CL83] M. Crandall and P.-L. Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

[CM90] A. Chorin and J. Marsden. *A Mathematical Introduction to Fluid Mechanics*. Springer, 1990.

[COO+18] 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.

[CS04] P. Cannarsa and C. Sinestrari. *Semiconcave Functions, Hamilton-Jacobi Equations, and Optimal Control*. Springer Science & Business Media, 2004.

[CS18] 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.

[DDB17] 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.

[DDC19] D. Davis, D. Drusvyatskiy, and V. Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

[DJ19] J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: A Hamiltonian perspective. *arXiv preprint arXiv:1906.00436*, 2019.

[DS01] J.-D. Deuschel and D. Stroock. *Large deviations*, volume 342. American Mathematical Society, 2001.

[Eva80] L. Evans. On solving certain nonlinear partial differential equations by accretive operator methods. *Israel Journal of Mathematics*, 36(3-4):225–247, 1980.

[Eva10] L. Evans. *Partial Differential Equations (Second Edition)*, volume 19. American Mathematical Society, 2010.

[Eva12] L. Evans. *An Introduction to Stochastic Differential Equations*, volume 82. American Mathematical Society, 2012.

[FW12] M. Freidlin and A. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260. Springer Science & Business Media, 2012.

[Gas07] S. Gasiorowicz. *Quantum Physics*. John Wiley & Sons, 2007.

[GLSS18] 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.

[HHS11] 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.

[HKN04] 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.

[HN05] B. Helffer and F. Nier. *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*. Springer, 2005.

[HRB08] E. Hazan, A. Rakhlin, and P. Bartlett. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems*, pages 65–72, 2008.

[HS12] P. Hislop and I. Sigal. *Introduction to Spectral Theory: With Applications to Schrödinger Operators*, volume 113. Springer Science & Business Media, 2012.

[HS20] H. He and W. J. Su. The local elasticity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

[Hwa80] C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

[HZRS16] 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.

[JGN+17] 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.

[JKA+17] 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.

[JKB+18] 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.

[Jor18] 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.

[KB14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KB17] W. Krichene and P. L. Bartlett. Acceleration and averaging in stochastic descent dynamics. In *Advances in Neural Information Processing Systems*, pages 6796–6806, 2017.

[KCD08] P. Kundu, I. Cohen, and D. Dowling. *Fluid Mechanics (Fourth Edition)*. Elsevier, 2008.

[KMN+16] 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.

[KP92] P. E. Kloeden and E. Platen. The approximation of multiple stochastic integrals. *Stochastic Analysis and Applications*, 10(4):431–441, 1992.

[Kri09] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

[Kro93] A. S. Kronfeld. Dynamics of Langevin simulations. *Progress of Theoretical Physics Supplement*, 111:293–311, 1993.

[KY03] H. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[LA19] Z. Li and S. Arora. An exponential learning rate schedule for deep learning. *arXiv preprint arXiv:1910.07454*, 2019.

[LH16] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[Lio82] P.-L. Lions. *Generalized Solutions of Hamilton-Jacobi Equations*, volume 69. London Pitman, 1982.

[LSJR16] 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.

[LTE17] 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.

[LWM19] 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, 2019.

[MHB16] S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 354–363, 2016.

[Mic19] L. Michel. About small eigenvalues of Witten Laplacian. *Pure and Applied Analysis*, 1(2), 2019.

[Mil75] G. N. Mil'shtein. Approximate integration of stochastic differential equations. *Theory of Probability & Its Applications*, 19(3):557–562, 1975.

[Mil86] G. N. Mil'shtein. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986.

[MV99] 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.

[Nie04] F. Nier. Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Journées Equations aux Dérivées Partielles*, pages 1–17, 2004.

[Pav14] G. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60. Springer, 2014.

[PT85] E. Pardoux and D. Talay. Discretization and simulation of stochastic differential equations. *Acta Applicandae Math*, 3:23–47, 1985.

[RRT17] 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.

[SBC16] 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.

[SDJS18] 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.

[SKYL17] 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.

[Smi17] 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.

[SS19] M. Sordello and W. J. Su. Robust learning rate selection for stochastic optimization via splitting diagnostic. *arXiv preprint arXiv:1910.08597*, 2019.

[SSG19] 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.

[Sun19] R. Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

[Tal82] D. Talay. *Analyse numérique des équations différentielles stochastiques*. PhD thesis, Université Aix-Marseille I, 1982.

[Tal84] 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.

[TH12] 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.

[Vil06] C. Villani. Hypocoercive diffusion operators. In *International Congress of Mathematicians*, volume 3, pages 473–498, 2006.

[Vil09] C. Villani. Hypocoercivity. *Memoirs of the American Mathematical Society*, 202(950), 2009.

[WWJ16] 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.

[YLWJ19] 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.

[ZBH+16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[Zei12] M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[ZLC17] 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.