COMMUN. MATH. SCI.
Vol. 18, No. 1, pp. 163–188

UNIFORM-IN-TIME WEAK ERROR ANALYSIS FOR STOCHASTIC GRADIENT DESCENT ALGORITHMS VIA DIFFUSION APPROXIMATION*

YUANYUAN FENG†, TINGRAN GAO‡, LEI LI§, JIAN-GUO LIU¶, AND YULONG LU‖

Abstract. Diffusion approximation provides weak approximation for stochastic gradient descent algorithms in a finite time horizon. In this paper, we introduce new tools motivated by the *backward error analysis* of numerical stochastic differential equations into the theoretical framework of diffusion approximation, extending the validity of the weak approximation from finite to infinite time horizon. The new techniques developed in this paper enable us to characterize the asymptotic behavior of constant-step-size SGD algorithms near a local minimum around which the objective functions are locally strongly convex, a goal previously unreachable within the diffusion approximation framework. Our analysis builds upon a truncated formal power expansion of the solution of a Kolmogorov equation arising from diffusion approximation, where the main technical ingredient is uniform-in-time bounds controlling the long-term behavior of the expansion coefficient functions near the local minimum. We expect these new techniques to bring new understanding of the behaviors of SGD near local minimum and greatly expand the range of applicability of diffusion approximation to cover wider and deeper aspects of stochastic optimization algorithms in data science.

Keywords. stochastic gradient descent; weak error analysis; diffusion approximation; stochastic differential equation; backward Kolmogorov equation.

AMS subject classifications. 60J20; 90C15.

1. Introduction

Stochastic gradient descent (SGD) is a prototypical stochastic optimization algorithm widely used for solving large scale data science problems [1–6], not only for its scalability to large datasets, but also due to its surprising capability of identifying parameters of deep neural network models with better generalization behavior than adaptive gradient methods [7–9]. The past decade has witnessed growing interests in accelerating this simple yet powerful optimization scheme [10–15], as well as better understanding its dynamics, through the lens of either discrete Markov chains [16, 17] or continuous stochastic differential equations [18–21].

This paper introduces new techniques into the theoretical framework of *diffusion approximation*, which provides *weak approximation* to SGD algorithms through the solution of a modified stochastic differential equation (SDE). Though numerous novel insights have been gained from this continuous perspective, it was previously unclear whether the modified SDEs can be adopted to study the asymptotic behavior of SGD, since the weak approximation is only valid over a finite time interval. In the nonconvex case, the approximation error blows up as time goes to infinity. For example, when the coefficient functions are bounded, the SDEs share the behaviors of random walks in

*Received: May 8, 2019; Accepted (in revised form): September 3, 2019. Communicated by Lexing Ying.

†Department of Mathematics, Penn State University, University Park, PA, 16802, USA (yzf58@psu.edu).

‡Committee on Computational and Applied Mathematics, Department of Statistics, University of Chicago, Chicago, IL, 60637, USA (tingrangao@galton.uchicago.edu).

§Corresponding Author. School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China (leili2010@sjtu.edu.cn).

¶Department of Mathematics and Department of Physics, Duke University, Durham, NC, 27708, USA (jliu@phy.duke.edu).

‖Department of Mathematics, Duke University, Durham, NC, 27708, USA (yulonglu@math.duke.edu).

high dimension space, which are transient. One will lose control of the system quickly as time goes on. In the strongly convex case, the problem remains open due to the unbounded diffusivity in the SDEs. We show in this paper that it is possible to study an *approximate solution* of the modified SDE for the latter case, which admits *uniform-in-time* weak error bounds and can thus be used for investigating the long-term behavior of SGD dynamics.

We concern ourselves in this paper with the problem of optimizing an empirical loss function $f:\mathbb{R}^d\to\mathbb{R}$

$$f(\theta)=\frac{1}{N_s}\sum_{i=1}^{N_s}\ell_\theta(z_i,y_i) \tag{1.1}$$

where $\{(z_i,y_i)\}_{i=1}^{N_s}$ are the training data (z_i's and y_i's are the data and labels, respectively) and $\ell_\theta(\cdot,\cdot)$ is the loss function with parameter θ to be learned. We will assume local strong convexity for f through the individual loss functions $\{\theta\mapsto\ell_\theta(z_i,y_i)\}_{i=1}^{N_s}$. The true gradient of f takes the form

$$\nabla f(\theta)=\frac{1}{N_s}\sum_{i=1}^{N_s}\nabla_\theta\ell_\theta(z_i,y_i). \tag{1.2}$$

The "stochastic gradient" considered in this paper are "mini-batches" subsampled from the summands $\{\nabla_\theta\ell_\theta(z_i,y_i)\}$ in (1.2), properly normalized so they provide an unbiased estimate for the true gradient. More specifically, fix a *batch size* parameter $B\in\mathbb{N}$, $1\le B\le N_s$, and let ξ be a subset of B distinct elements uniformly sampled from the integers $\{1,\dots,N_s\}$ without replacement, we set

$$\nabla f(\theta;\xi):=\frac{1}{B}\sum_{j\in\xi}\nabla_\theta\ell_\theta(z_j,y_j). \tag{1.3}$$

Such constructed stochastic gradients are unbiased estimates of the true gradient in the sense that $\mathbb{E}_\xi[\nabla f(\cdot;\xi)]=\nabla f$.

Below, we will use x to mean the parameter θ and X_n to mean the discrete iterates in SGD, as is standard in numerical analysis of SDEs. The notation "$\mathbb{E}_x$" will be used to mean expectation under the initial condition $X(0)=x$ for SDE or $X_0=x$ for the SGD iterates. Also, Ξ will be used to denote the set of all possible values of ξ, and in the situation described above, it is the set of all subsets of $\{1,2,\dots,N_s\}$ with size B. The iterative stochastic numerical scheme under consideration throughout this paper is

$$X_{n+1}=X_n-\eta\nabla f(X_n;\xi_n),\quad n=0,1,\dots \tag{1.4}$$

where $\eta>0$ is the constant step size and $\nabla f(\cdot;\xi_n)$ is the stochastic gradient with $\xi_n\in\Xi$ being i.i.d.. We characterize the asymptotic distributional behavior of the iterates $\{X_n\}_{n\ge0}$ as n approaches infinity, by adapting tools from *backward error analysis* of stochastic numerical schemes [22–27] to modified SDEs arising from the diffusion approximation [18–20]. So far, asymptotic analyses for the dynamics of (1.4) have been made possible only through the Markov chain techniques [16, 17]. We also refer to [28, 29] for some convergence analysis of stochastic gradient descent methods for continuous time models. This paper is our first attempt at fully unleashing the rich and powerful SDE techniques for studying stochastic numerical optimization schemes in large scale statistical and machine learning.

1.1. Main contribution: Long-time weak approximation for SGD via SDE. The dynamics of discrete, iterative numerical algorithms can often be better understood from their continuous time limit, typically described by ordinary differential equations. This perspective has been proven fruitful in the analysis of many deterministic optimization algorithms [30–34]. An analogy of this type of continuous-time-limit analysis for SGD algorithms is provided by the *diffusion approximation* [18,20]: in any *finite* time interval, the distribution of X_n defined by the SGD dynamics (1.4) is close to the distribution of the solution of the following SDE at time $t=n\eta$:

$$\mathrm{d}X=-\nabla\left[f(X)+\frac{1}{4}\eta\|\nabla f(X)\|^2\right]\mathrm{d}t+\sqrt{\eta\Sigma(X)}\mathrm{d}W, \tag{1.5}$$

where

$$\Sigma=\mathbb{E}_\xi\left[(\nabla f(\cdot;\xi)-\nabla f)\otimes(\nabla f(\cdot;\xi)-\nabla f)\right]$$

is the covariance matrix of the random gradients, and W is the standard Brownian motion [35]. In numerical SDE literature, SDE of type (1.5) is often referred to as the *stochastic modified equations*; they play an important role in constructing high-order numerical approximation schemes for invariant measures of ergodic SDEs (see, e.g., [24,25]). In the context of data science, diffusion approximation has been used to gain insights into online PCA [20], entropy-SGD [36,37], and nonconvex optimization [21], to name just a few.

Despite its effectiveness as a continuous analogy of stochastic numerical optimization algorithms, the range of applicability of diffusion approximation is significantly limited by its restricted validity in a finite time interval. In particular, this means that the solution of the SDE (1.5) can be used to rigorously approximate only a finite number (though very large) of SGD iterates (1.4), and thus can not be used in the same way as Markov-chain-based theoretical analysis [16,38,39] to study the asymptotic behavior of $\{X_n\}_{n\geq 0}$ as $n\to\infty$. This paper aims at closing this theoretical gap by extending the validity of diffusion approximation from finite- to infinite-time horizon. To the best of our knowledge, this is the first work that studies the asymptotic distributional behavior of SGD from an SDE perspective.

Our main technical contribution in this paper is to adopt the framework of *weak backward error analysis* to the solution $u=u(x,t)=\mathbb{E}_x[\varphi(X(t))]$ of the following backward Kolmogorov equation associated with SDE (1.5):

$$\begin{aligned}&\frac{\partial u}{\partial t}=-\nabla f\cdot\nabla u+\eta\left(-\frac{1}{4}\nabla\|\nabla f\|^2\cdot\nabla u+\frac{1}{2}\mathrm{Tr}\left(\Sigma\nabla^2 u\right)\right)\\&u(x,0)=\varphi(x)\end{aligned} \tag{1.6}$$

where we recall that $\mathbb{E}_x$ stands for taking expectation under the initial condition $X(0)=x$, $\mathrm{Tr}(A)$ stands for the trace of a square matrix A, $\Sigma=\Sigma(x)$ is the covariance matrix as in (1.5), and ∇u, $\nabla^2 u$ denote the gradient and Hessian of $u=u(x,t)$ with respect to the spatial variable x. The function $\varphi:\mathbb{R}^d\to\mathbb{R}$ is an arbitrary "observable" of the stochastic dynamical system that characterizes properties of interest of the iterates $\{X_n\}_{n\geq 0}$. Weak error analysis concerns the behavior of $\{\varphi(X_n)\}_{n\geq 0}$ for any φ with sufficient regularity; for instance, by taking $\varphi=f$, we can study the asymptotic oscillatory and/or concentration behavior of the objective values $f(X_n)$ with respect to the global minimum if standard convexity assumptions are imposed on f.

In a nutshell, backward error analysis is based on identifying the associated generator of a numerical scheme with the generator of a modified SDE, up to higher order

terms in the powers of the step size η. This can be achieved, e.g., by formally expanding the generator of the modified SDE into a power series of the step size, and then determining the coefficients (which are functions of the space and time variables, but not the step size η) of this power series using information from the numerical scheme; it is then natural to expect that a proper truncation of this formal power series can be used as a reasonable approximation for the iterates of the stochastic numerical scheme (in the weak sense), even though the formal series may not converge (and thus the solution of the SDE may not be a good approximation for the discrete iterates for all time). As illustrated by many examples in the numerical analysis of ergodic SDEs (see, e.g., [22–27] and the references therein), it turns out that the coefficient functions of the formal power series capture—in a uniform-in-time fashion—the leading order behavior of the discrete numerical scheme; this enables practitioners to draw conclusion on the closeness between the invariant measure of the numerical scheme and the invariant measure of the truncated formal series. In other words, though solutions of (1.6) can not be used directly to capture the long-term behavior of SGD (1.4), we construct an alternative, auxiliary function approximation of the solution of (1.6), which turns out to be a superior weak approximation of (1.4) in the sense that the approximation error is uniform-in-time and in higher powers of the step size η. The time-uniformity of such a truncated formal series approximation enables us to study the asymptotic distributional behavior of the iterates of (1.4), thus closing the gap in the theoretical analysis between diffusion approximation and Markov-chain-based analysis. We provide an overview for the main steps in our analysis in the next section.

1.2. Sketch of the main approach. We consider a formal expansion of the solution $u = u(x,t) = \mathbb{E}_x[\varphi(X(t))]$ of (1.6) in a power series with respect to the step size $\eta > 0$:

$$u(x,t) = \sum_{\ell=0}^{\infty} \eta^\ell u_\ell(x,t). \tag{1.7}$$

For the ease of exposition, let us introduce short-hand notations $\mathcal{L}_1, \mathcal{L}_2$ for the differential operators appearing in the right-hand side of (1.6):

$$\mathcal{L}_1 := -\nabla f \cdot \nabla, \qquad \mathcal{L}_2 := -\frac{1}{4}\nabla \|\nabla f\|^2 \cdot \nabla + \frac{1}{2}\mathrm{Tr}\left(\Sigma \nabla^2\right) \tag{1.8}$$

with which (1.6) can be recast into

$$\begin{aligned} &\partial_t u = \mathcal{L}_1 u + \eta \mathcal{L}_2 u, \\ &u(x,0) = \varphi(x). \end{aligned} \tag{1.9}$$

Formally plugging (1.7) into (1.9) and equating terms corresponding to the same powers of η, we can determine all coefficient functions $u_n(x,t)$ from solving corresponding PDEs, namely, for $\ell = 0$

$$\begin{aligned} &\partial_t u_0 = \mathcal{L}_1 u_0, \\ &u_0(x,0) = \varphi(x) \end{aligned} \tag{1.10}$$

and for $\ell \geq 1$

$$\begin{aligned} &\partial_t u_\ell = \mathcal{L}_1 u_\ell + \mathcal{L}_2 u_{\ell-1}, \\ &u_\ell(x,0) = 0. \end{aligned} \tag{1.11}$$

Determining any u_ℓ can thus be done by inductively solving a sequence of first-order PDEs (1.10)-(1.11). In fact, with some work we can establish exponential convergence of each u_ℓ to its equilibrium state as t approaches infinity, provided that f is strongly convex.

We then construct an approximation for u by truncating the formal series (1.7), yielding

$$u^N(x,t)=\sum_{\ell=0}^{N}\eta^\ell u_\ell(x,t). \tag{1.12}$$

If the formal series (1.7) converges uniformly, u^N is certainly a good approximation of u up to an order $\mathcal{O}\left(\eta^{N+1}\right)$ error. The crux of our argument is that, even when the convergence of (1.7) is not guaranteed, it turns out that we can still use $\left\{u^1(x,n\eta)\right\}_{n\geq 0}$ as good approximation for $\{\mathbb{E}_x[\varphi(X_n)]\}_{n\geq 0}$ (recall that $\mathbb{E}_x$ represents the expectation conditioned on the initial condition $X_0=x$); most notably, the $\mathcal{O}\left(\eta^2\right)$ error in this approximation is bounded uniformly in n, allowing us to draw quantitative conclusions on the asymptotic distributional behavior of $\mathbb{E}_x[\varphi(X_n)]$ from that of $u^1(x,n\eta)$. Since u^1 corresponds to a measure ν^1 independent of the test function φ, our argument then justifies that the measure ν^1 approximates the distribution of the SGD with second order weak accuracy. It is very tempting to push this idea further by considering u^N, $N>1$ in place of u^1 and expecting it to better approximate $\mathbb{E}_x[\varphi(X_n)]$ up to higher orders of error; however, our analysis indicates that in general $\left|u^N(x,n\eta)-\mathbb{E}_x[\varphi(X_n)]\right|=\mathcal{O}\left(\eta^2\right)$ can no longer be improved by choosing $N>1$, even though u^N could be a better approximation for the solution u of the backward Kolmogorov equation (1.6) when $N>1$.

The superior, uniform-in-time approximation of the truncated formal expansion to $\mathbb{E}_x[\varphi(X_n)]$ is achieved by the fact that the coefficient functions u_ℓ are totally determined by the local behavior of f and φ (i.e. behaviors on compact sets), whereas the solution u of (1.6) depends on the global information and is thus harder to control. Due to this locality, the local strong convexity of f then leads to the exponential decay of the derivatives for the coefficient functions u_ℓ, which finally gives the uniform-in-time weak approximation. This will become transparent after we establish Theorem 2.1. The locality can be illustrated by a toy SDE example in one dimension with $f(x)=\frac{1}{2}x^2$, and $\Sigma(x)\equiv 1$. Note that this SDE example is simply given to illustrate the roles of u_ℓ and why they are local, while it is not necessarily the diffusion approximation of some SGD iteration. In this example, SDE (1.5) corresponds to an Ornstein–Uhlenbeck process, and the solution of (1.6) adopts the explicit integral representation

$$\begin{aligned} u(x,t)&=\frac{1}{\sqrt{2\pi S}}\int_{\mathbb{R}^d}\varphi(w)\exp\left(-\frac{(w-xe^{-(1+2\eta)t})^2}{2S}\right)\mathrm{d}w\\ &=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}^d}\varphi\left(xe^{-(1+2\eta)t}+\sqrt{S}y\right)\exp(-y^2/2)\,\mathrm{d}y \end{aligned} \tag{1.13}$$

where

$$S=\frac{\eta}{2(1+2\eta)}\left(1-e^{-2(1+2\eta)t}\right).$$

We can obtain a formal expansion of $u(x,t)$ in terms of η using a Taylor expansion for φ at $xe^{-(1+2\eta)t}$ in the integrand of (1.13). We keep $2m$ terms in the Taylor expansion and note that all odd powers of $\sqrt{S}$ vanish, which leads to the following expansion of

error $\mathcal{O}\left(\eta^{m+1}\right)$:

$$u(x,t)=\sum_{k=0}^{m}\frac{1}{(2k)!}\varphi^{(2k)}\left(xe^{-(1+2\eta)t}\right)S^k\cdot\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}^d}y^{2k}\exp(-y^2/2)\,dy+O\left(\eta^{m+1}\right).$$

The u_ℓ's can then be obtained by further expanding the functions about $\eta=0$ and combining terms of equal powers. Clearly, such obtained u_ℓ's in this expansion will only depend on the derivatives of φ at xe^{-t}; meaning that $u_\ell(x,t)$ only depends on the behaviors of φ inside the ball with radius $|x|$, whereas for any x, $u(x,t)$ depends on the values of φ in the whole space. The formal series expansion is like the Taylor series of the function $u(x,t)$ with respect to η. As known, in general one can not expect the Taylor series to converge to the original function unless the function is analytic, which exactly resembles the difference between the solution of (1.6) and the truncated formal series expansion (1.12): the latter maintains only the barely minimum local information in the diffusion approximation for characterizing the asymptotic distributional behavior of the dynamics of SGD (1.4).

Full details of our theoretical framework can be found in Section 2 and the appendices.

1.3. Outline. In the remainder of this paper, we present our main theorems and main proofs in Section 2, and validate our theory with numerical experiments in Section 3. Technical lemmas and auxiliary results are deferred to the appendices. We conclude this paper and propose future directions in Section 4.

2. Main results

We begin by stating the assumption that will be used throughout this paper (recall that Ξ is the set of all possible values of the random parameters ξ).

ASSUMPTION 2.1. *Without loss of generality, assume f has a local minimum at the origin $x_*=0$. Gradients of the random functions $\left\{f(\cdot;\xi)\in C^3(\mathbb{R}^d)\,|\,\xi\in\Xi\right\}$ provide unbiased estimates for the gradient of f, i.e., $\mathbb{E}_\xi[\nabla f(x;\xi)]=\nabla f(x)$ for all $x\in\mathbb{R}^d$. Moreover, we assume the following hold for the random functions. There exists $R_1>0$ such that*

(1) Each random function $f(\cdot;\xi)$ is γ-strongly convex in $B(x_*,R_1)$, i.e., $f(\cdot;\xi)-\frac{1}{2}\gamma\|\cdot\|^2$ is convex for all $\xi\in\Xi$;

(2) The random gradients at $x_*=0$ are bounded:

$$\sup_\xi\|\nabla f(0;\xi)\|\le b<\infty. \tag{2.1}$$

for some $b>0$ and more over

$$R_1>\frac{16b}{3\gamma}=:R_0. \tag{2.2}$$

Though our assumption on the individual $f(\cdot;\xi)$'s appears to be strong, it is not particularly restrictive for the most commonly encountered scenario of SGD application where each random function $f(\cdot;\xi)$ is constructed from the same loss function $loss(y_\xi,g(z_\xi))\equiv\ell_\theta(z_\xi,y_\xi)$, and the only source of randomness is in the random data (z_ξ,y_ξ) sampled from an unknown data distribution. In this case, Assumption 2.1 can be stated just once for the loss function, as done in [40]. Such an assumption on the individual summands in the empirical loss function has also appeared previously in

Markov-chain-based studies of stochastic gradient descent algorithms, e.g. Assumption A4 in [16]. The boundedness assumption (2.1) is obviously satisfied if the loss function $\ell_\theta(z_i,y_i)$ is bounded at $\theta=0$ for all data (z_i,y_i).

In the remainder of this section, we divide our exposition of the main results into two parts. Estimates establishing the exponential convergence of the coefficient functions of the formal series expansion appear in Section 2.1, and their applications to studying the asymptotic distributional behavior of SGD iterates appear in Section 2.2.

2.1. Formal series expansion. Under the local strong convexity assumption in Assumption 2.1, the following two lemmas can be easily established. We defer the proofs to Appendix A. In particular, the convergence in Wasserstein-2 distance in Lemma 2.2 is well-known (see, e.g., Proposition 1 in [16]); we include a simple proof in Appendix A for completeness. In the rest of this paper, for any $R>0$, we denote $B(0,R)$ for the Euclidean ball of radius R centered at the origin (which is also the global minimum of f by Assumption 2.1).

LEMMA 2.1. *Suppose Assumption 2.1 holds, and denote* $R_0=16b/3\gamma$. *If* $R\in(R_0,R_1]$, *set*

$$\eta_0=\min\left\{\frac{1}{2\gamma},\frac{3R}{8b},\frac{3\gamma R^2/8-2bR}{2\gamma bR+b^2}\right\}.$$

Then for any $\eta\le\eta_0$ *and* $X_0=x\in B(0,R)$, *we have* $X_n\in B(0,R)$ *for all* $n\ge0$. *In other words, under these assumptions the sequence generated by the SGD is uniformly bounded in both* n *and* ξ.

LEMMA 2.2. *Suppose Assumption 2.1 holds, and let* μ_n *denote the law of the* n^{th} *iterate* X_n *of SGD (1.4). Assume* $\operatorname{supp}\mu_0\subset B(0,R)$ *with* $R\in(R_0,R_1]$ *and denote* $L=\sup_\xi\sup_{\|x\|\le R}\|\nabla^2 f(x;\xi)\|$, *where* $\|\nabla^2 f(x;\xi)\|$ *is the spectral norm (largest singular value) of the Hessian matrix* $\nabla^2 f(x;\xi)$. *Then, when* η *is sufficiently small,* μ_n *converges to a probability measure* π *under the Wasserstein-2 norm (*W_2*-norm) at exponential rate*

$$W_2(\mu_n,\pi)\le C\rho^n$$

for $\rho=(1-2\gamma\eta+\eta^2L)^{1/2}$.

REMARK 2.1. Clearly, for different local minima around which the loss functions are locally strongly convex, the probability measure π will be different. Since the SDEs in diffusion approximation have a nonzero transition probability connecting any two points in space, the diffusion approximation cannot be uniform in time for such *globally nonconvex* cases. Even for such globally nonconvex loss functions, our theory indicates that the local information of diffusion approximation is enough to capture the long-time behavior of SGD near the local minimum. To obtain global diffusion approximation for such nonconvex cases, one has to modify the values of the loss function outside the region where SGD can see.

We define the $S^{(n)}$ operator by

$$(S^{(n)}\varphi)(x):=\mathbb{E}_x[\varphi(X_n)]=\int_{\mathbb{R}^d}\varphi(y)\,\mu_n(\mathrm{d}y). \tag{2.3}$$

Fixing any smooth test function φ, we denote

$$U^n(x):=S^{(n)}\varphi(x). \tag{2.4}$$

We know from [20] that S is L^∞-nonexpansive, and that $\{S^{(n)}\}$ is a semi-group generated by S such that

$$S^{(n)} = S^n := S \circ S \ldots \circ S \ (n \text{ copies}). \tag{2.5}$$

Since convergence in Wasserstein distance implies weak convergence, Lemma 2.2 implies

$$\lim_{n\to\infty} U^n = \int_{\mathbb{R}^d} \varphi \mathrm{d}\pi. \tag{2.6}$$

However, this does not provide much precise and/or quantitative information regarding how U^n converges to $\int_{\mathbb{R}^d} \varphi \mathrm{d}\pi$. An important goal of this paper is to shed new lights on the dynamics of μ_n as $n \to \infty$. Within the diffusion approximation framework, it can be shown (see, e.g., [20]) that the semi-group evolution U^n admits a *weak second order diffusion approximation* over a finite time interval $[0,T]$, in the sense that for all sufficiently smooth φ there holds

$$\sup_{n \le T/\eta} \|U^n(\cdot) - u(\cdot, n\eta)\|_{L^\infty} \le C(T,\varphi,\eta_0)\eta^2 \tag{2.7}$$

for all $\eta \le \eta_0$, where $\eta_0 > 0$ is a constant, and $u(x,t) = \mathbb{E}_x[\varphi(X(t))]$ is the solution of the backward Kolmogorov Equation (1.6). Roughly speaking, SDE (1.5) can be regarded as the weak approximation of the SGD (1.4) over any finite time interval $[0,T]$. Unfortunately, the validity of this approximation for infinite time ($T \to \infty$) is unclear. For nonconvex objective functions, it is known that the approximation can break down quickly as $T \to \infty$. One obvious example is the situation described in Remark 2.1. For globally and strongly convex objective functions (which generate confining dynamics for SGD, according to Lemma 2.1), the validity of long-time diffusion approximation is still in doubt due to the unboundedness diffusivity encoded in Σ. As motivated in Section 1.2, we will switch gears and use a truncated formal series (1.12) in place of the solution u of (1.6) to approximate U^n, for all arbitrarily large $n \ge 0$.

Before stating the main technical result concerning the exponential convergence of the u_ℓ's in the formal asymptotic expansion, we introduce another notation to simplify the exposition and proof: denote

$$I_k = \{J = (j_1, j_2, \ldots, j_k) : 1 \le j_k \le d\}.$$

For $J \in I_k$, we denote

$$\partial^J u := \partial_{j_1} \ldots \partial_{j_k} u.$$

We write $J_0 \le J$ if $\partial^J u$ is a partial derivative of $\partial^{J_0} u$, and $J_1 = J - J_0$ if $\partial^J = \partial^{J_0}\partial^{J_1}$.

REMARK 2.2. The reason that we adopt the notation ∂^J instead of the standard multi-index notation ∂^α where $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $\alpha_1 + \ldots + \alpha_d = k$ is mainly for the sake of clarity and simplicity of exposition. First, this convention is widely used for tensor analysis in physics and engineering. More importantly, in Appendix B where we prove Theorem 2.1, $\sum_{J \in I_{n+1}} \partial_t (\partial^J u)^2$ naturally has a quadratic form associated with the Hessian matrix $\nabla^2 f$ so that we can make use of the strong convexity. If we use ∂^α notation, we will have to multiply some weight factors w_α such that $\sum_{|\alpha|=k} w_\alpha \partial_t (\partial^\alpha u)^2$ has the desired quadratic form.

We are now ready to present our main estimates for the exponential rate of decay for the coefficient functions in the formal series expansion (1.7). We will use P to denote a generic polynomial whose concrete form may change from line to line. The number of arguments for the polynomials will also be clear in the context, which we will not emphasize.

THEOREM 2.1. *Assume Assumption 2.1 holds, $\eta \le \eta_0$ and $R \in (R_0, R_1]$, for $R_0 > 0$, $R_1 > 0$ and $\eta_0 > 0$ defined as in Lemma 2.1. Recall that $x^* = 0$ is the unique minimum of f.*

(i) For an arbitrary test function $\varphi \in C^1(\mathbb{R}^d)$, u_0 satisfies

$$\sup_{x \in B(0,R)} |u_0(x,t) - \varphi(0)| \le R \|\varphi\|_{C^1(B(0,R))} e^{-\gamma t}. \tag{2.8}$$

In addition, if $\varphi \in C^k(B(0,R))$ and $f \in C^{k+1}(B(0,R))$ for some integer $k \ge 1$, then

$$\sup_{J \in I_k} \sup_{x \in B(0,R)} |\partial^J u_0(x,t)| \le P\Big(\|\varphi\|_{C^k(B(0,R))}, \|f\|_{C^{k+1}(B(0,R))} \Big) e^{-\gamma t}. \tag{2.9}$$

(ii) For any $n \ge 1$, if the test function $\varphi \in C^{2n+1}(B(0,R))$ and $f \in C^{2n+2}(B(0,R))$, then for any $\gamma' < \gamma$,

$$\begin{aligned} &\sup_{x \in B(0,R)} |u_n(x,t) - \varphi_n| \\ &\le P\Big(\|\varphi\|_{C^{2n+1}(B(0,R))}, \|f\|_{C^{2n+2}(B(0,R))}, \|\Sigma\|_{C^{2n-1}(B(0,R))} \Big) e^{-\gamma' t}, \end{aligned} \tag{2.10}$$

where

$$\varphi_n := \int_0^\infty \mathcal{L}_2 u_{n-1}(0,s)\, ds. \tag{2.11}$$

In addition, if $\varphi \in C^{k+2n}(B(0,R))$ and $f \in C^{k+1+2n}(B(0,R))$ for some $k \ge 1$, then for any $\gamma'' < \gamma$,

$$\begin{aligned} &\sup_{J \in I_k} \sup_{x \in B(0,R)} |\partial^J u_n(x,t)| \\ &\le P\Big(\|\varphi\|_{C^{k+2n}(B(0,R))}, \|f\|_{C^{k+1+2n}(B(0,R))}, \|\Sigma\|_{C^{k+2n-2}(B(0,R))} \Big) e^{-\gamma'' t}. \end{aligned} \tag{2.12}$$

The proof of Theorem 2.1 is quite technical; we defer full details to Appendix B. We state an immediate corollary of Theorem 2.1 to close this subsection.

COROLLARY 2.1. *Under the same assumptions as in Theorem 2.1, the truncated formal series u^N defined in (1.12) "approximately satisfies" the backward Equation (1.6) in the sense that*

$$\partial_t u^N = (\mathcal{L}_1 + \eta \mathcal{L}_2) u^N - \eta^{N+1} \mathcal{L}_2 u_N. \tag{2.13}$$

Consequently, if $\varphi \in C^{2k+2N}(B(0,R))$ and $f \in C^{2k+1+2N}(B(0,R))$ for some $k \ge 1$, we have

$$\sup_{x \in B(0,R)} |\partial_t^k u^N - (\mathcal{L}_1 + \eta \mathcal{L}_2)^k u^N| \le C(N,R) e^{-\gamma t/2} \eta^{N+1} \tag{2.14}$$

where $C(N,R)=Q_{N,k}\left(\|\varphi\|_{C^{2k+2N}(B(0,R))},\|f\|_{C^{2k+1+2N}(B(0,R))},\|\Sigma\|_{C^{2k}(B(0,R))}\right)$ *for some polynomial* $Q_{N,k}$.

It is clear from Theorem 2.1 that all the coefficient functions $u_n(x,t)$ depend only on the information of f and Σ inside the ball $B(0,\|x\|)$, in the sense that the bound does not change if we modify the values of φ, f, and a outside $B(0,\|x\|)$. Thus u_n reflects the "local information" of u. This is in stark contrast with the solution of (1.6) at x, which inevitably depends on the values of φ outside $B(0,\|x\|)$ due to the parabolicity of the second order PDE (1.6). As explained in Section 1.2, this is due to the fact that the $u_n(x,t)$'s are essentially the "Taylor expansion coefficients" of u with respect to the step size. This is also the reason that we referred to (1.7) as only a formal series expansion: in general the Taylor series need not converge to the original function. See also the Ornstein–Uhlenbeck process example in Section 1.2 for a concrete example.

2.2. Dynamics of SGD with constant step size. In this subsection we apply the results from Section 2.1 to study the asymptotic distributional behavior of the SGD dynamics (1.4). Throughout the rest of this subsection, we always assume that $X_0\in B(0,R)$ and R satisfies the condition of Lemma 2.1. The confining nature of the dynamics allows us to choose very general functions as test functions, e.g., smooth functions that grow exponentially as $\|x\|\to\infty$, for the weak approximation results to hold. This is because we can always modify the part of the test function outside of $B(0,R)$. More precisely, we have

LEMMA 2.3. *Under Assumption 2.1, given any test function* $\varphi\in C^k(\mathbb{R}^d)$ *for some* $k\in\mathbb{N}$, *we can choose* $\tilde{\varphi}\in C^k\left(\mathbb{R}^d\right)$ *compactly supported such that*

$$\|\tilde{\varphi}\|_{C^k(\mathbb{R}^d)}\le C\|\varphi\|_{C^k(B(0,R))}$$

and

$$\mathbb{E}_x\left[\varphi(X_n)\right]=\mathbb{E}_x\left[\tilde{\varphi}(X_n)\right],\ \forall x\in B(0,R).$$

Similarly, in the formal series expansion (1.7) *for the diffusion approximation, replacing* φ *with* $\tilde{\varphi}$ *does not change any of the coefficient functions* $u_\ell(x,t),x\in B(0,R),\ell\ge 0$.

Lemma 2.3 is a simple consequence of transport Equations (1.10)-(1.11). Notably, we emphasize again that the locality of the coefficient functions $u_\ell(x,t)$ is in stark contrast with the solution of the backward Kolmogorov Equation (1.6), since (1.6) has diffusion effects which are global. Lemma 2.3 indicates we can focus on test functions compactly supported near the local minimum we care about. The main result of this paper is the following.

THEOREM 2.2. *Assume Assumption 2.1 holds,* $\eta\le\eta_0$ *and* $R\in(R_0,R_1]$, *for* $R_0>0$, $R_1>0$ *and* $\eta_0>0$ *defined as in Lemma 2.1. If* $f(\cdot;\xi)\in C^7(B(0,R))$ *and* $\varphi\in C^6(B(0,R))$, *then* $u^1=u_0+\eta u_1$ *approximates the dynamics of SGD* (1.4) with weak second order, *in the sense that there exists a positive constant* $C(\varphi,f,R)$ *independent of* n *such that*

$$\sup_{x\in B(0,R)}\left|\mathbb{E}_x\left[\varphi(X_n)\right]-u^1(x,n\eta)\right|\le C(\varphi,f,R)\eta^2. \tag{2.15}$$

Proof. By Lemma 2.3, we can assume without loss of generality that φ is compactly supported and $\|\varphi\|_{C^k(\mathbb{R}^d)}\le C_k\|\varphi\|_{C^k(B(0,R))}$ for sufficiently large k. Let us recall

the notation $U^n(x)=\mathbb{E}_x[\varphi(X_n)]$ and that $S:L^\infty\left(\mathbb{R}^d\right)\to L^\infty\left(\mathbb{R}^d\right)$ forms the semi-group $S^{(n)}=S^n$. Thus,

$$U^{n+1}(x)=\mathbb{E}(U^n(x-\eta\nabla f(x;\xi)))\coloneqq SU^n(x).$$

Noticing that $U^n(x)=S^n\varphi(x)$ and $\varphi(x)=u^N(x,0)$, by a telescoping sum we have

$$U^n(x)-u^N(x,n\eta)=\sum_{j=1}^{n}S^{n-j}(Su^N(x,(j-1)\eta)-u^N(x,j\eta)).$$

By the fact that S is L^∞ nonexpansive,

$$|U^n(x)-u^N(x,n\eta)|\le\sum_{j=1}^{n}\|Su^N(x,(j-1)\eta)-u^N(x,j\eta)\|_{L^\infty}. \tag{2.16}$$

We fix $N=1$ and for the sake of convenience, we introduce

$$t_j\coloneqq j\eta. \tag{2.17}$$

By Corollary 2.1, it holds for $t\in[t_{j-1},t_j]$ that

$$u^1(x,t)=u^1(x,t_{j-1})+\int_{t_{j-1}}^{t}(\mathcal{L}_1+\eta\mathcal{L}_2)u^1(x,s)\,ds-\eta^2\int_{t_{j-1}}^{t}\mathcal{L}_2u_1(x,s)\,ds. \tag{2.18}$$

Substituting this expression of u^1 into the right-hand side (and repeatedly for some terms), one has

$$\begin{aligned}u^1(x,t)=&\,u^1(x,t_{j-1})+(t-t^n)\mathcal{L}_1u^1(x,t_{j-1})+\eta(t-t^n)\mathcal{L}_2u^1(x,t_{j-1})\\&+\frac{1}{2}(t-t^n)^2\mathcal{L}_1^2u^1(x,t_{j-1})+\eta\int_{t_{j-1}}^{t}\int_{t_{j-1}}^{s}(\mathcal{L}_2(\mathcal{L}_1+\eta\mathcal{L}_2)+\mathcal{L}_1\mathcal{L}_2)u^1(x,\tau)\,d\tau ds\\&+\int_{t_{j-1}}^{t}\int_{t_{j-1}}^{s}\int_{t_{j-1}}^{\tau}\mathcal{L}_1^2(\mathcal{L}_1+\eta\mathcal{L}_2)u^1\,dzd\tau ds-\eta^2\int_{t_{j-1}}^{t}\mathcal{L}_2u_1\,ds\\&-\eta^2\int_{t_{j-1}}^{t}\int_{t_{j-1}}^{s}(\mathcal{L}_1+\eta\mathcal{L}_2)\mathcal{L}_2u_1\,d\tau ds-\eta^2\int_{t_{j-1}}^{t}\int_{t_{j-1}}^{s}\int_{t_{j-1}}^{\tau}\mathcal{L}_1^2\mathcal{L}_2u_1\,dzd\tau ds.\end{aligned} \tag{2.19}$$

Hence,

$$\begin{aligned}&\left|u^1(x,j\eta)-u^1(x,(j-1)\eta)-\eta(\mathcal{L}_1+\eta\mathcal{L}_2)u^1(x,(j-1)\eta)-\frac{\eta^2}{2}\mathcal{L}_1^2u^1(x,(j-1)\eta)\right|\\&\le C\eta^3\sup_{t\in[t_{j-1},t_j]}\left(\sum_{I=1}^{4}\sup_{x\in B(0,R)}(|\partial^Iu^1|+|\partial^Iu_1|)\right).\end{aligned} \tag{2.20}$$

By Theorem 2.1,

$$\begin{aligned}&\left|u^1(x,j\eta)-u^1(x,(j-1)\eta)-\eta(\mathcal{L}_1+\eta\mathcal{L}_2)u^1(x,(j-1)\eta)-\frac{\eta^2}{2}\mathcal{L}_1^2u^1(x,(j-1)\eta)\right|\\&\le C(f,\varphi,R)\eta^3e^{-\gamma(j-1)\eta/2}.\end{aligned} \tag{2.21}$$

In the meanwhile, applying Taylor expansion to $Su^1(x,(j-1)\eta)=\mathbb{E}\left[u^1(x-\eta\nabla f(x,\xi)),(j-1)\eta\right]$ and applying Theorem 2.1 gives

$$\left|Su^1(x,(j-1)\eta)-u^1(x,(j-1)\eta)-\eta(\mathcal{L}_1+\eta\mathcal{L}_2)u^1(x,(j-1)\eta)-\frac{\eta^2}{2}\mathcal{L}_1u^1(x,(j-1)\eta)\right|$$
$$\leq C(f,\varphi,R)\eta^3e^{-\gamma(j-1)\eta/2}. \tag{2.22}$$

Combining (2.21) and (2.22), we have

$$|Su^1(x,(j-1)\eta)-u^1(x,j\eta)|\leq C(f,\varphi,R)\eta^3e^{-\gamma(j-1)\eta/2}$$

and thus the right-hand size of (2.16) can be further bounded by

$$|U^n(x)-u^1(x,n\eta)|\leq C(f,\varphi,R)\eta^2$$

for some positive constant $C(f,\varphi,R)$ independent of n. This completes the proof. □

The key contribution of Theorem 2.2 is the extension of the range of applicability of diffusion approximation (2.7) from finite time interval $[0,T]$ to infinite time. A direct consequence is the following description of the "weak expansion" of the stationary distribution of the dynamics (1.4).

COROLLARY 2.2. *Under the same conditions as in Theorem 2.2, we have for all* $n\gtrsim\frac{1}{\eta}\log(1/\eta)$ *that*

$$\sup_{x\in B(0,R)}|\mathbb{E}_x\varphi(X_n)-\varphi(0)|=\sup_{x\in B(0,R)}|U^n(x)-\varphi(0)|\leq C(\varphi,f,R)\eta$$

for some positive constant $C(\varphi,f,R)$. *Moreover, the probability measure in Lemma 2.2 satisfies*

$$\left|\int_{\mathbb{R}^d}\varphi\,\mathrm{d}\pi-\varphi(0)-\eta\varphi_1\right|\leq C\eta^2,$$

where $\varphi_1=\lim_{t\to\infty}u_1(x,t)=\int_0^\infty\mathcal{L}_2u_0(0,s)\,ds$ *is independent of* x.

The conclusion follows immediately from noting that, for $n\gtrsim\eta^{-1}\log(\eta^{-1})$,

$$|u^1(x,n\eta)-\varphi(0)|=|u_0(x,n\eta)+\eta u_1(x,n\eta)-\varphi(0)|\leq C(f,\varphi,R)\eta.$$

In particular, if we choose $\varphi=f$, Corollary 2.2 tells us that SGD descends the value of a strongly convex objective function to an $\mathcal{O}(\eta)$ neighborhood of the global minimum in only $\mathcal{O}\left(\eta^{-1}\log(\eta^{-1})\right)$ time. Measured in the time scale of diffusion approximation, where $t=n\eta$ in $u^1(x,n\eta)$, this is equivalent to saying that the SGD dynamics reduces the objective value to $\mathcal{O}(\eta)$ away from the global minimum within time $n\eta=\mathcal{O}(\log(1/\eta))$, which is exponentially fast, as well known.

At last, we remark that if X_0 starts with a measure μ_0 instead of $X_0=x$, then $\int(u_0+\eta u_1)(t,x)\mu_0(dx)$ will approximate $\mathbb{E}\varphi(X_n)$ uniformly in time. We may further rewrite the quantity as

$$\int_{\mathbb{R}^d}(u_0(x,t)+\eta u_1(x,t))\mu_0(dx)=\int_{\mathbb{R}^d}\varphi(x)(\nu_0(dx)+\eta\nu_1(dx)), \tag{2.23}$$

with ν_0,ν_1 respectively satisfying (see Appendix C for a formal derivation):

$$\partial_t\nu_0-\nabla\cdot(\nabla f\nu_0)=0,\quad \nu_0(0)=\mu_0, \tag{2.24}$$

and

$$\partial_t \nu_1 - \nabla\cdot(\nabla f \nu_1) = \frac{1}{4}\nabla\cdot(\nabla\|\nabla f\|^2\nu_0) + \frac{1}{2}\partial_{ij}(\Sigma_{ij}\nu_0), \quad \nu_1(0) = 0. \tag{2.25}$$

Theorem 2.2 then implies that $\nu^1 := \nu_0 + \eta\nu_1$ or $\frac{\nu_0+\eta\nu_1}{M(\nu_0+\eta\nu_1)}$ approximates the distribution of X_n with second weak order, where $M(\nu)$ means the total mass of ν:

$$M(\nu) := \int_{\mathbb{R}^d} d\nu.$$

Remark 2.3. The weak order of approximation $\mathcal{O}(\eta^2)$ in Theorem 2.2 is optimal in the sense that no higher order approximation error can be achieved by choosing $N > 1$ in (1.12), although the formal truncated series u^N may better approximate the Kolmogorov Equation (1.6). This is because the diffusion approximation itself is only a weak second order approximation for SGD [20, Theorem 2.2]. Higher order approximation for the SGD dynamics requires higher derivatives of u in the PDE (1.6), but it no longer describes a diffusion process (solutions of Itô equations).

3. Numerical experiments

In this section we demonstrate the approximation power of the truncated formal series (1.12) with numerical experiments for some one-dimensional ($d = 1$) examples. We consider SGD schemes

$$f(x;\xi) = f(x) + \frac{\xi}{2}x, \quad x \in \mathbb{R} \tag{3.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is locally strongly convex near one of its local minima, and ξ is a Rademacher random variable that assigns equal probability $1/2$ to both -1 and $+1$. Following the definitions in (1.8), we have explicitly

$$\mathcal{L}_1 = -f'(x)\partial_x, \qquad \mathcal{L}_2 = -\frac{1}{2}f'(x)f''(x)\partial_x + \frac{1}{8}\partial_x^2. \tag{3.2}$$

The first two terms in the formal series expansion (1.7) can be determined by solving the two first order PDEs sequentially: First solve

$$\begin{aligned} &\partial_t u_0 + f'(x)\partial_x u_0(x,t) = 0 \\ &u_0(x,0) = \varphi(x) \end{aligned} \tag{3.3}$$

to get

$$u_0(x,t) = \varphi(x_0(x,t)), \tag{3.4}$$

where $x_0(x,t)$ is the intercept of the characteristic line passing through $(x,t) \in \mathbb{R}\times\mathbb{R}_+$. We then use (3.4) to solve

$$\begin{aligned} &\partial_t u_1 + f'(x)\partial_x u_1(x,t) = -\frac{1}{2}f'(x)f''(x)\partial_x u_0 + \frac{1}{8}\partial_x^2 u_0 \\ &u_1(x,0) = 0 \end{aligned} \tag{3.5}$$

which gives

$$u_1(x,t) = -\frac{1}{2}f'(x_0(x,t))\varphi'(x_0(x,t))\log\frac{f'(x)}{f'(x_0(x,t))}$$

$$+\frac{1}{8}f'(x_0(x,t))f''(x_0(x,t))\varphi'(x_0(x,t))\int_{x_0(x,t)}^{x}\frac{\mathrm{d}\xi}{[f'(\xi)]^3}$$
$$-\frac{1}{16}f'(x_0(x,t))\varphi'(x_0(x,t))\left\{\frac{1}{[f'(x_0(x,t))]^2}-\frac{1}{[f'(x)]^2}\right\}$$
$$+\frac{1}{8}[f'(x_0(x,t))]^2\varphi''(x_0(x,t))\int_{x_0(x,t)}^{x}\frac{\mathrm{d}\xi}{[f'(\xi)]^3}. \tag{3.6}$$

Details of this computation can be found in Appendix D.

EXAMPLE 3.1. *We consider a simple example*

$$f(x)=\frac{1}{2}x^2-\frac{1}{2}x. \tag{3.7}$$

The stochastic gradient updates are

$$X_{n+1}=X_n-\eta\nabla f(X_n;\xi_n)=(1-\eta)X_n-\frac{\eta}{2}(1-\xi_n)$$

where $\{\xi_n\}_{n\geq 0}$ *are i.i.d. standard Rademacher random variables. The limiting distribution of this Markov chain is identical to that of* $X_\infty=\eta\sum_{j=0}^{\infty}\theta_j(1-\eta)^j$ *where the* θ_j*'s are i.i.d.* Bernoulli$(1/2)$ *random variables. The infinite series converges whenever* $\eta\in(0,1)$*, but the stationary distribution is drastically different for different values of* η *[41, §2.5]: If* $\eta=1/2$*,* X_∞ *is uniformly distributed on* $[0,1]$*; if* $1/2<\eta<1$*, the distribution of* X_∞ *is singular (supported on a set of Lebesgue measure* 0*); if* $0<\eta<1/2$*, for some values of* η *the stationary distribution is singular, but it has also been established that for almost all* $\eta\in(0,1/2)$ *the stationary distribution is absolutely continuous. We are most interested in the regime* $\eta\in(0,1/2)$ *where* η *is small.*

We choose several different test functions φ *to verify the order of the weak approximation error between* $U^n(x)=\mathbb{E}_x[\varphi(X_n)]$ *and* $u^1=u_0+\eta u_1$ *established in Theorem 2.2. The results are summarized in Figure 3.1 and Figure 3.2.*

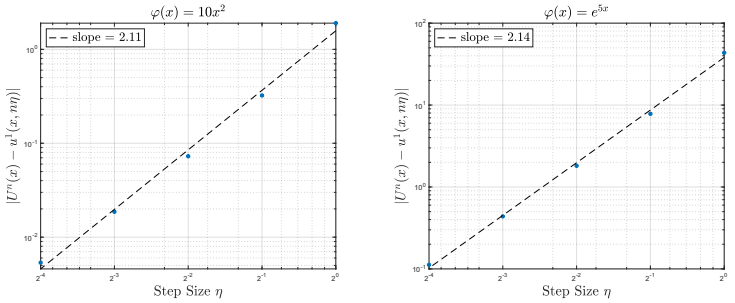

FIG. 3.1. *Log-log plots numerically verifying the weak second order diffusion approximation established in Theorem 2.2, using example (3.7) and two different test functions* φ*. For each* φ*, we fix* $x=1$ *and* $n\eta=5$*, then let* η *vary in* $\{2^{-4},2^{-3},2^{-2},2^{-1},2^0\}$*. We use a Monte–Carlo simulation to evaluate* $U^n(x)=\mathbb{E}_x[\varphi(X_n)]$*, by averaging* $\varphi(X_n)$ *over* 10^8 *independent trajectories starting from* $X_0=x$*. The slopes of the fitting lines are close to* 2*, which justify the second order approximation established in Theorem 2.2.*

EXAMPLE 3.2. *We now consider a more complicated example in which the gradient* ∇f *is nonlinear. Set*

$$f(x)=\frac{1}{2}x^2+0.1x^3 \tag{3.8}$$

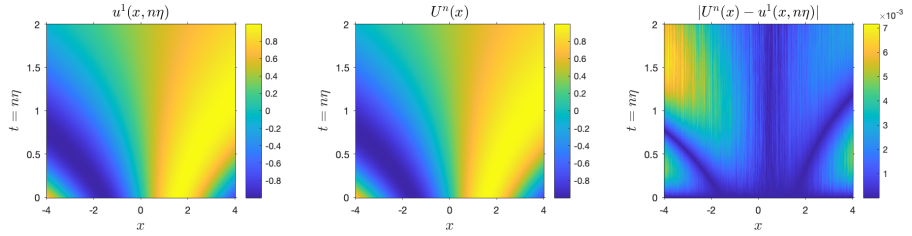

FIG. 3.2. *Visual comparison of* $u^1(x,n\eta)$ *and* $U^n(x)$ *for* $\varphi(x)=\sin(x)$ *over* $(x,t)\in[-4,4]\times[0,2]$, *with* $\eta=0.01$. *Each* $U^n(x)$ *is evaluated over* 10^4 *independent trajectories generated from the gradient dynamics associated with* (3.7).

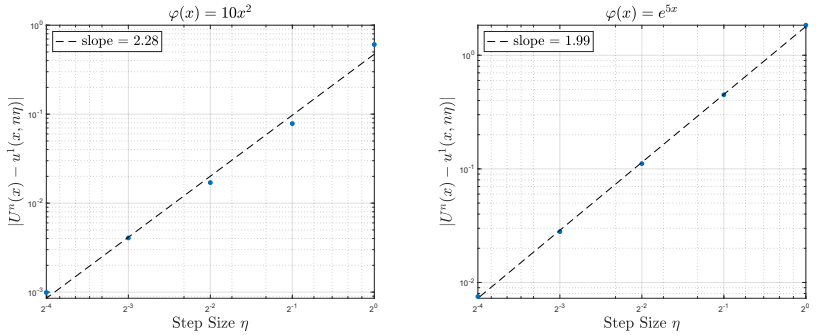

FIG. 3.3. *Log-log plots numerically verifying the weak second order diffusion approximation established in Theorem 2.2, using example* (3.8) *and two different test functions* φ. *For each* φ, *we fix* $x=1$ *and* $n\eta=5$, *then let* η *vary in* $\{2^{-4},2^{-3},2^{-2},2^{-1},2^0\}$. *We use a Monte–Carlo simulation to evaluate* $U^n(x)=\mathbb{E}_x[\varphi(X_n)]$, *by averaging* $\varphi(X_n)$ *over* 10^8 *independent trajectories starting from* $X_0=x$. *The slopes of the fitting lines are close to* 2, *which justify the second order approximation established in Theorem 2.2.*

and the stochastic gradient updates can be written as

$$X_{n+1}=X_n-\eta\nabla f(X_n;\xi_n)=(1-\eta)X_n-0.3\eta X_n^2-\frac{\eta}{2}\xi$$

where $\{\xi_n\}_{n\geq 0}$ *are i.i.d. standard Rademacher random variables. We choose the same test functions* φ *as in Example 3.1. The results are summarized in Figure 3.3.*

4. Conclusion

In this paper, we establish uniform-in-time weak error bounds for diffusion approximation of SGD algorithms, under the local strong convexity assumption for the objective functions. To this end, we adapted the idea of backward error analysis in numerical SDEs, and used a truncated formal series expansion with respect to the constant step size for the backward Kolmogorov equation associated with the modified SDE—instead of the solution itself—to approximate the SGD iterates for arbitrarily long time. This enables us to draw quantitative conclusions for the weak asymptotic behavior of the SGD iterates from estimates of the coefficient functions of the truncated formal expansion, which is the first result of this type for diffusion-approximation-based SGD analysis. We believe the tools developed in this paper have great potential in generalizing the range of applicability of diffusion approximation to many other stochastic optimization algorithms in data science, such as SGD with non-constant step size and momentum-based acceleration techniques.

Acknowledgement. The work of T. Gao was partially supported by NSF DMS-1854831. The work of L. Li was partially sponsored by NSFC 11901389, Shanghai Sailing Program 19YF1421300 and NSFC 11971314. The work of J.-G. Liu was partially supported by KI-Net NSF RNMS11-07444 and NSF DMS-1812573.

Appendix A. Proofs of technical lemmas in Section 2.1.

Proof. **(Proof of Lemma 2.1.)** By (1.4), we have

$$\begin{aligned}\|X_{n+1}\|^2 &= \|X_n\|^2 - 2\eta(\nabla f(X_n;\xi)-\nabla f(0;\xi))\cdot X_n + \eta^2\|\nabla f(X_n;\xi)\|^2 - 2\eta\nabla f(0;\xi)\cdot X_n \\ &\le \|X_n\|^2 - 2\gamma\eta\|X_n\|^2 + \eta^2(b+\gamma|X_n|)^2 - 2\eta\nabla f(0;\xi)\cdot X_n, \end{aligned} \tag{A.1}$$

where we applied the strong convexity of $f(\cdot;\xi)$ in the last inequality and the fact that

$$|\nabla f(X_n;\xi_n)| \le |\nabla f(0;\xi_n)| + \gamma|X_n| \le b+\gamma|X_n|.$$

When $\|X_n\| \le \frac{R}{2}$, (A.1) can be further controlled by

$$\|X_{n+1}\|^2 \le (1-2\gamma\eta)\frac{R^2}{4} + \eta^2 b^2 + \eta^2\gamma bR + \eta^2\frac{\gamma^2R^2}{4} + \eta bR.$$

Noting that $-\frac{1}{2}\gamma\eta R^2 + \eta^2\frac{\gamma^2R^2}{4} \le -\frac{3}{8}\gamma\eta R^2$, we find

$$\begin{aligned}\|X_{n+1}\|^2 &\le \frac{R^2}{4} + \gamma\eta R(-\frac{3R}{8}+\eta b) + \eta b(\eta b + R) \\ &\le \frac{R^2}{4} + \frac{3R}{8} * \frac{11R}{8} < R^2.\end{aligned}$$

When $\frac{R}{2} \le |X_n| \le R$, we have

$$\begin{aligned}\|X_{n+1}\|^2 &\le \|X_n\|^2 + (-2\gamma\eta+\eta^2\gamma^2)\cdot\frac{R^2}{4} + 2\gamma bR\eta^2 + \eta^2b^2 + 2\eta bR \\ &\le \|X_n\|^2 + (2b - \frac{3\gamma R^2}{8})\eta + \eta^2(2\gamma bR + b^2) \le |X_n|^2.\end{aligned}$$

Thus the conclusion follows. ∎

Proof. **(Proof of Lemma 2.2.)** Consider two copies of the chain

$$Y_{n+1} = Y_n - \eta\nabla f(Y_n;\xi_n), \quad Z_{n+1} = Z_n - \eta\nabla f(Z_n;\xi_n). \tag{A.2}$$

The two chains are coupled through the random variable ξ_n. This means that they pick the same function to compute the gradient at every iteration n. Meanwhile, each chain has the same asymptotic distributional behavior as the SGD. We then have

$$\begin{aligned}\mathbb{E}\|Y_{n+1}-Z_{n+1}\|^2 = \mathbb{E}\|Y_n-Z_n\|^2 &- 2\eta\mathbb{E}[(Y_n-Z_n)\cdot(\nabla f(Y_n,\xi_n)-\nabla f(Z_n,\xi_n))] \\ &+ \eta^2\mathbb{E}\|\nabla f(Y_n,\xi_n)-\nabla f(Z_n,\xi_n)\|^2.\end{aligned}$$

For the second term, we use conditional expectation to deduce that

$$\begin{aligned}&\mathbb{E}[(Y_n-Z_n)\cdot(\nabla f(Y_n,\xi_n)-\nabla f(Z_n,\xi_n))] \\ =&\mathbb{E}[(Y_n-Z_n)\cdot\mathbb{E}[\nabla f(Y_n,\xi_n)-\nabla f(Z_n,\xi_n)|Y_m,Z_m,m\le n]] \\ =&\mathbb{E}[(Y_n-Z_n)\cdot(\nabla f(Y_n)-\nabla f(Z_n))] \ge \gamma\mathbb{E}\|Y_n-Z_n\|^2.\end{aligned}$$

The last term is upper bounded by

$$\eta^2\mathbb{E}\|\nabla f(Y_n,\xi_n)-\nabla f(Z_n,\xi_n)\|^2\le\eta^2L^2\mathbb{E}|Y_n-Z_n|^2.$$

Therefore, it follows that

$$\mathbb{E}\|Y_{n+1}-Z_{n+1}\|^2\le(1-2\gamma\eta+\eta^2L^2)\mathbb{E}\|Y_n-Z_n\|^2.$$

Now, if $\eta<2\gamma/L^2$, then $0<1-2\gamma\eta+\eta^2L^2<1$. We claim that under this choice of η, the law of X_n is a Cauchy sequence under the W_2 norm. In fact, for any $\epsilon>0$, we can pick $m>0$ such that $(2R)^2(1-2\gamma\eta+\eta^2L^2)^m<\epsilon^2/4$. For $n\ge m$, we pick Y_0 to have the same distribution as X_0 and Z_0 to have the same distribution as X_{n-m}. Then, Y_m has the same distribution as X_m while Z_m has the same distribution as X_n. Moreover,

$$\mathbb{E}\|Y_m-Z_m\|^2\le(1-2\gamma\eta+\eta^2L^2)^m\mathbb{E}\|Y_0-Z_0\|^2<\epsilon^2/4. \tag{A.3}$$

It follows that

$$\left(\mathbb{E}\|Y_m-Z_m\|^2\right)^{1/2}<\epsilon/2.$$

We recall that the Wasserstein-2 distance is given by

$$W_2(\mu,\nu)=\left(\inf_{\gamma\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^2d\gamma\right)^{1/2}, \tag{A.4}$$

where $\Pi(\mu,\nu)$ means the set of all the joint distributions γ whose marginal distributions are μ and ν respectively. Since the joint distribution of (Y_m,Z_m) is in $\Pi(\mu_n,\mu_m)$, one finds $W_2(\mu_n,\mu_m)<\epsilon/2$. This means that μ_n is a Cauchy sequence, and it holds for some probability distribution π that

$$\lim_{n\to\infty}W_2(\mu_n,\pi)=0.$$

Finally, we obtain from (A.3) that

$$W_2(\mu_n,\mu_m)\le(1-2\gamma\eta+\eta^2L^2)^{m/2}\sqrt{\mathbb{E}\|Y_0-Z_0\|^2}\le C(1-2\gamma\eta+\eta^2L)^{m/2},$$

where C is independent of m,n (the second moment of X_{n-m} is uniformly bounded). The conclusion follows from taking the limit $n\to\infty$. ∎

Appendix B. Proof of the exponential decay estimates.

Proof. **(Proof of Theorem 2.1.)** The genesis of the exponential decay rates of the u_ℓ's can be traced back to the following simple yet important observation: Suppose $y(t)$ satisfies

$$\dot{y}=-\nabla f(y) \tag{B.1}$$

with $y(0)=x$, then $\|y(t)\|$ is a non-increasing function and

$$\|y(t)\|\le\|y(0)\|e^{-\gamma t}. \tag{B.2}$$

We now begin our proof. First by the method of characteristics [42, Theorem 5.34], one notices that u_0 satisfies

$$\partial_t u_0+\nabla f(x)\cdot\nabla u_0=0, \tag{B.3}$$

$$u_0(x,0)=\varphi(x). \tag{B.4}$$

Let y be the function in (B.1) with $y(0)=x\in B(0,R)$. And for any given $T>0$, $t\in[0,T]$, define $z(t):=y(T-t)$. Then it follows that

$$u_0(z(t),t)=\varphi(z(0)),\ \forall t\in[0,T].$$

Consequently, we have $u_0(x,t)=\varphi(y(t)),\ \forall t>0$. Hence,

$$|u_0(x,t)-\varphi(0)|\le\|\nabla\varphi\|_{L^\infty(B(0,R))}|y(t)|\le R\|\varphi\|_{C^1(B(0,R))}e^{-\gamma t}.$$

For the estimate of derivatives, we use induction. When $k=1$, following from Equations (B.3) and (B.4), we have

$$\partial_t\|\nabla u_0\|^2=-2\nabla u_0\cdot\nabla^2 f\cdot\nabla u_0-\nabla f\cdot\nabla\|\nabla u_0\|^2\quad\text{and}\quad\|\nabla u_0(x,0)\|^2=\|\nabla\phi(x)\|^2.$$

Since f is strongly convex,

$$\partial_t\|\nabla u_0\|^2\le-2\gamma\|\nabla u_0\|^2-\nabla f\cdot\nabla\|\nabla u_0\|^2. \tag{B.5}$$

Recall $y(t)$ which was defined in equation (B.1) and $z(t)=y(T-t)$. By chain rule, equation (B.5) yields that $\frac{d}{dt}\|\nabla u_0(z(t),t)\|^2\le-2\gamma\|\nabla u_0(z(t),t)\|^2$, which by Grönwall's inequality further yields

$$\|\nabla u_0(z(t),t)\|\le e^{-\gamma t}\|\nabla u_0(z(0),0)\|\le e^{-\gamma t}\|\nabla\varphi(z(0))\|,\ \forall t\in[0,T].$$

This then yields

$$\begin{aligned}\|\nabla u_0(x,t)\|&\le e^{-\gamma t}\|\nabla\varphi(y(t))\|\le\|\nabla\varphi\|_{L^\infty(B(0,R))}e^{-\gamma t}\\&\le\|\varphi\|_{C^1(B(0,R))}e^{-\gamma t},\ \forall t>0,x\in B(0,R).\end{aligned}$$

Hence inequality (2.9) is verified for $k=1$. By induction, we assume for any $k\le m$, inequality (2.9) holds. Next we study the case for $k=m+1$. For $J\in I_{m+1}$, we differentiate equation (B.3) by ∂^J and get $\partial_t\partial^J u_0+\partial^J(\nabla f\cdot\nabla u_0)=0$. Then multiplying both sides by $\partial^J u_0$ and summing over all $J\in I_{m+1}$ gives

$$\partial_t v=-2\sum_{J\in I_{m+1}}\sum_{i=1}^d\partial^J u_0\partial^J(\partial_i f\partial_i u_0),$$

where $v=\sum_{J\in I_{m+1}}(\partial^J u_0)^2$. We note that the right-hand side can be splitted into the sum of three terms according to the general Leibniz rule in calculus. And then the above equation becomes

$$\partial_t v\le-2(n+1)\gamma v-\nabla f\cdot\nabla v-2\sum_{i=1}^d\sum_{J\in I_{m+1}}\partial^J u_0\sum_{J_0\le J,|J_0|\ge2}\partial^{J_0}\partial_i f\partial^{J-J_0}\partial_i u_0. \tag{B.6}$$

Here is a brief explanation of the above inequality (B.6). For $k\in\{1,\cdots,m+1\}$, $j_k\in\{1,\cdots,d\}$, putting the first order derivative ∂_{j_k} on $\partial_i f$ and $\partial^{J-\{j_k\}}$ on $\partial_i u_0$, we would obtain

$$-2\sum_{J\in I_{m+1}}\sum_{k=1}^{m+1}\sum_{i,j_k=1}^d\partial^J u_0\partial_{j_k}\partial_i f\partial^{J-\{j_k\}}\partial_i u_0,$$

which is a quadratic form associated with the Hessian matrix $\nabla^2 f$. This also explains why we do not use the traditional definition of multi-index in our paper (the question related to Remark 2.2). By the strong convexity of f, the above term is bounded above by

$$-2\sum_{J\in I_{m+1}}\sum_{k=1}^{m+1}\sum_{i,j_k=1}^{d}\gamma(\partial^{J-\{j_k\}}\partial_i u_0)^2,$$

which can be further bounded above by $-2(m+1)\gamma v$. Putting all the J^{th} derivatives on $\partial_i u_0$ yields the second term $-\nabla f\cdot\nabla v$. For the third term, we only need to consider the remaining terms due to the Leibniz rule. Hence the validity of (B.6) has been proved.

For the last term in (B.6), we use Young's inequality and the induction assumption, then derive that

$$\partial_t v \le -2m\gamma v - \nabla f\cdot\nabla v + P\Big(\|\varphi\|_{C^m(B(0,R))}, \|f\|_{C^{m+2}(B(0,R))}\Big)e^{-2\gamma t}.$$

We also note that $|v(x,0)|\le\|\varphi\|_{C^{m+1}(B(0,R))}$, for $x\in B(0,R)$. Hence we get

$$v(z(t),t)\le P\Big(\|\varphi\|_{C^{m+1}(B(0,R))}, \|f\|_{C^{m+2}(B(0,R))}\Big)e^{-2\gamma t},\ \forall t\in[0,T].$$

This then gives

$$v(x,t)\le P\Big(\|\varphi\|_{C^{m+1}(B(0,R))}, \|f\|_{C^{m+2}(B(0,R))}\Big)e^{-2\gamma t},\ \forall t>0.$$

Hence result (2.9) is proved.

Now we start to study u_n. The equation which u_n satisfies is the following

$$\begin{aligned}&\partial_t u_n + \nabla f\cdot\nabla u_n = \mathcal{L}_2 u_{n-1},\\ &u_n(x,0)=0.\end{aligned}$$

Based on this, we could write down a formula for u_n,

$$u_n(x,t)=\int_0^t \mathcal{L}_2 u_{n-1}(y(s),t-s)\,ds. \tag{B.7}$$

Here we recall that y satisfies Equations (B.1) with $y(0)=x\in B(0,R)$ and thus (B.2).

Consider $n=1$. For convenience, we denote

$$g(x,t)=\mathcal{L}_2 u_0(x,t).$$

Intuitively, the limiting behavior of $u_1(x,t)$ is determined by $g(0,t)$. We now verify this.

Recalling the definition of the operator $\mathcal{L}_2$ (1.8), we have

$$\begin{aligned}\sup_{x\in B(0,R)}(|g(x,t)|+|\nabla g(x,t)|)&\le C\Big(\|f\|_{C^3(B(0,R))}+\|\Sigma\|_{C^1(B(0,R))}\Big)\|u_0\|_{C^3(B(0,R))}\\ &\le P\Big(\|\varphi\|_{C^3(B(0,R))}, \|f\|_{C^4(B(0,R))}, \|\Sigma\|_{C^1(B(0,R))}\Big)e^{-\gamma t},\end{aligned} \tag{B.8}$$

where the last inequality followed from (2.9). It follows that $\sup_{x\in B(0,R)}|u_1(x,t)|$ is uniformly bounded in t. Moreover, we further split u_1 as

$$u_1(x,t)=\int_0^t g(0,t-s)\,ds+\int_0^t (g(y(s),t-s)-g(0,t-s))\,ds. \tag{B.9}$$

The second term is controlled directly by (B.8) as

$$\begin{aligned}\left|\int_0^t (g(y(s),t-s)-g(0,t-s))\,ds\right| &\le \int_0^t \|\nabla g(\cdot,t-s)\|_{L^\infty(B(0,R))}|y(s)|\,ds\\ &\le C\int_0^t e^{-\gamma(t-s)}e^{-\gamma s}\,ds\\ &= Cte^{-\gamma t}\le Ce^{-\gamma' t},\end{aligned}$$

where $C=P\Big(\|\varphi\|_{C^3(B(0,R))},\|f\|_{C^4(B(0,R))},\|\Sigma\|_{C^1(B(0,R))}\Big)$ for some polynomial P and the last inequality followed from $te^{-\gamma t}\le C(\gamma')e^{-\gamma' t}$ for any $\gamma'<\gamma$.

Regarding the first term in (B.8), we know that it converges to

$$\varphi_1=\int_0^\infty g(0,s)\,ds,$$

with the exponential rate. Hence, overall, we have

$$|u_1(x,t)-\varphi_1|\le P\Big(\|\varphi\|_{C^3(B(0,R))},\|f\|_{C^4(B(0,R))},\|\Sigma\|_{C^1(B(0,R))}\Big)e^{-\gamma' t}. \tag{B.10}$$

For the derivatives of u_1, we notice that

$$\partial_t\|\nabla u_1\|^2=-2\nabla u_1\cdot\nabla^2 f\cdot\nabla u_1-\nabla f\cdot\nabla\|\nabla u_1\|^2+2\sum_{j=1}^d\partial_j u_1\partial_j(\mathcal{L}_2u_0). \tag{B.11}$$

Also we notice that

$$\sup_{x\in B(0,R)}|\partial_j(\mathcal{L}_2u_0(x,t))|\le C\Big(\|f\|_{C^3(B(0,R))}+\|\Sigma\|_{C^1(B(0,R))}\Big)\|u_0\|_{C^3(B(0,R))}.$$

We use this in (B.11), and for the first term we use strong convexity of f as well, then get

$$\begin{aligned}\partial_t\|\nabla u_1\|^2\le&-2\gamma\|\nabla u_1\|^2-\nabla f\cdot\nabla\|\nabla u_1\|^2\\ &+2\sum_{j=1}^d|\partial_j u_1|\ P\Big(\|\varphi\|_{C^3(B(0,R))},\|f\|_{C^4(B(0,R))},\|\Sigma\|_{C^1(B(0,R))}\Big)e^{-\gamma t}.\end{aligned}$$

We then apply Young's inequality to further get that, for any $\gamma''<\gamma$, there exists a polynomial P in $\|\varphi\|_{C^3(B(0,R))}$, $\|f\|_{C^4(B(0,R))}$ and $\|\Sigma\|_{C^1(B(0,R))}$ such that

$$\begin{aligned}\partial_t\|\nabla u_1\|^2\le&-2\gamma''\|\nabla u_1\|^2-\nabla f\cdot\nabla\|\nabla u_1\|^2\\ &+P\Big(\|\varphi\|_{C^3(B(0,R))},\|f\|_{C^4(B(0,R))},\|\Sigma\|_{C^1(B(0,R))}\Big)e^{-2\gamma t}.\end{aligned}$$

Hence it holds that

$$\sup_{x\in B(0,R)}\|\nabla u_1(x,t)\|\le P\Big(\|\varphi\|_{C^3(B(0,R))},\|f\|_{C^4},\|\Sigma\|_{C^1(B(0,R))}\Big)e^{-\gamma'' t}. \tag{B.12}$$

For the higher derivatives of u_1, the analysis goes similarly as that of u_0. We also use induction here. Assume for any $k\le m$, (2.12) holds. For $k=m+1$, we denote $w=\sum_{J\in I_{m+1}}(\partial^J u_1)^2$ and get

$$\partial_t w\le-2(m+1)\gamma w-\nabla f\cdot\nabla w-2\sum_{i=1}^d\sum_{J\in I_{m+1}}\partial^J u_1\sum_{J_0\le J,|J_0|\ge 2}\partial^{J_0}\partial_i f\partial^{J-J_0}\partial_i u_1$$

$$+2\sum_{J\in I_{m+1}}(\partial^J u_1)\partial^J\Big(\mathcal{L}_2 u_0\Big)$$
$$\leq -2\gamma w-\nabla f\cdot\nabla w+P\Big(\|\varphi\|_{C^{m+3}(B(0,R))},\|f\|_{C^{m+4}(B(0,R))},\|\Sigma\|_{C^{m+1}B(0,R)}\Big)e^{-2\gamma'' t}.$$

From this we get

$$\sup_{x\in B(0,R)} w(x,t)\leq P\Big(\|\varphi\|_{C^{m+3}(B(0,R))},\|f\|_{C^{m+4}(B(0,R))},\|\Sigma\|_{C^{m+1}(B(0,R))}\Big)e^{-2\gamma'' t}.$$

This shows that (2.12) is true for $n=1$, $k=m+1$. Hence (2.12) holds for all derivatives of u_1.

The analysis for $n\geq 2$ is similar to the case $n=1$ and can be performed using induction. This completes the proof. ∎

Appendix C. Formal derivation of the equations of the measures. In this section, we aim to derive the Equations (2.24)-(2.25) in a formal way. Observe that ν_0 is a probability measure so the equation of ν_0 can be derived from the empirical measure $\frac{1}{N}\sum_i\delta(x-X_i(t))$ where each X_i satisfies the transport Equation (B.1). However, this cannot be generalized to the equation of ν_1. Hence we adopt another different formal derivation as follows.

According to $u(x,t)=\mathbb{E}_x\varphi(X(t))$, we expect u_0 to be written as

$$u_0(x,t)=\int_{\mathbb{R}^d}\varphi(y)G_0(dy,t;x).$$

According to the definition of ν_0, one has

$$\nu_0(\cdot,t)=\int_{\mathbb{R}^d}G_0(\cdot,t;x)\mu_0(dx),$$

and thus $G_0(dy,t;x)$ means the Green's function for the evolution of ν_0 with initial condition $X(0)=x$, or $\delta(\cdot-x)$. By the equation of $u_0(x,t)$, it is easy to find that G_0 satisfies

$$\partial_t G_0(\cdot,t;x)+\nabla f(x)\cdot\nabla_x G_0(\cdot,t;x)=0. \tag{C.1}$$

Due to the Markovian property of the dynamics, we can easily infer that the measure ν_0 satisfies

$$\nu_0(\cdot,t)=\int_{\mathbb{R}^d}G_0(\cdot,t-s;y)\nu_0(dy;s)=:\mathcal{I}^{(0)}_{t-s}\nu_0(\cdot;s). \tag{C.2}$$

Here, $\mathcal{I}^{(0)}_{t-s}$ is the evolution operator. Using (C.1), one finds

$$\partial_t\int_{\mathbb{R}^d}G_0(\cdot,t-s;x)\nu_0(dx;s)+\int_{\mathbb{R}^d}\nu_0(dx;s)\nabla f(x)\cdot\nabla_x\mathcal{I}^{(0)}_{t-s}\delta(\cdot-x)=0$$

or

$$\partial_t\nu_0(\cdot,t)-\int_{\mathbb{R}^d}\nabla_x\cdot\Big(\nabla f(x)\nu_0(dx;s)\Big)\mathcal{I}^{(0)}_{t-s}\delta(\cdot-x)=0.$$

Since $\mathcal{I}^{(0)}_{t-s}$ is independent of x, the second term is then reduced to

$$-\int_{\mathbb{R}^d}\nabla_x\cdot\Big(\nabla f(x)\nu_0(dx;s)\Big)\mathcal{I}^{(0)}_{t-s}\delta(\cdot-x)=-\mathcal{I}^{(0)}_{t-s}\int_{\mathbb{R}^d}\nabla_x\cdot\Big(\nabla f(x)\nu_0(dx;s)\Big)\delta(\cdot-x)$$
$$=-\mathcal{I}^{(0)}_{t-s}\nabla\cdot(\nabla f\nu_0(\cdot;s)).$$

Taking $t \to s$, one obtains the equation for ν_0.

Similarly, let $G_1(\cdot,t;x)$ satisfy the following inhomogeneous equation

$$\begin{aligned}&\partial_t G_1(\cdot,t;x) + \nabla f(x)\cdot\nabla_x G_1(\cdot,t;x) = -\frac{1}{4}\nabla\|\nabla f\|^2\cdot\nabla_x G_0(\cdot,t;x) + \frac{1}{2}\mathrm{Tr}(\Sigma\nabla_x^2 G_0(\cdot,t;x)),\\ &G_1(\cdot,0;x) = 0.\end{aligned} \tag{C.3}$$

Then, we have

$$u_1(x,t) = \int_{\mathbb{R}^d} \varphi(y) G_1(y,t;x)dy, \tag{C.4}$$

and

$$\nu_1(\cdot,t) = \int_{\mathbb{R}^d} G_1(\cdot,t;x)\mu_0(dx). \tag{C.5}$$

By the linearity, one has

$$\nu_1(\cdot,t) = \int_{\mathbb{R}^d} G_1(\cdot,t-s;x)\nu_0(dx,s) + \mathcal{I}^{(0)}_{t-s}\nu_1(\cdot,s). \tag{C.6}$$

The first term arises from (C.5) with zero initial data while the second term is from the homogeneous part with initial data $\nu_1(\cdot,s)$. Setting $t \to t-s$ in (C.3), multiplying $\nu_0(dx,s)$ and integrating, one has

$$\begin{aligned}&\partial_t \int_{\mathbb{R}^d} G_1(\cdot,t-s;x)\nu_0(dx,s) - \int_{\mathbb{R}^d} G_1(\cdot,t-s;x)\nabla\cdot(\nabla f(x)\nu_0(dx,s))\\ =&\mathcal{I}^{(0)}_{t-s}\left(\frac{1}{4}\nabla\cdot(\nabla\|\nabla f\|^2\nu_0) + \frac{1}{2}\partial_{ij}(\Sigma_{ij}\nu_0)\right).\end{aligned}$$

Clearly, the second term $\mathcal{I}^{(0)}_{t-s}\nu_1(\cdot,s)$ satisfies

$$\partial_t \mathcal{I}^{(0)}_{t-s}\nu_1(\cdot,s) - \nabla\cdot(\nabla f\mathcal{I}^{(0)}_{t-s}\nu_1(\cdot,s)) = 0.$$

Adding up the above two equations and taking $t \to s$ yields

$$\partial_t\nu_1 - \nabla\cdot(\nabla f\nu_1) = \frac{1}{4}\nabla\cdot(\nabla\|\nabla f\|^2\nu_0) + \frac{1}{2}\partial_{ij}(\Sigma_{ij}\nu_0).$$

REMARK C.1. The generalization to ν_n for $n \geq 2$ is more involved and the equation for ν_n is similar to ν_1. The key relation is some analogue of (C.6), given by $\nu_n(\cdot,t) = \sum_{m=0}^n \int G_m(\cdot,t-s;x)\nu_{n-m}(dx,s)$ due to linearity. (In fact, one may also expand the Fokker-Planck equation for the diffusion approximation in terms of η to obtain the equations for ν_n. However, this type of derivation does not give the inisight into the dynamics.)

Appendix D. Computations for the numerical examples. In this appendix we include detailed computations used in the numerical examples in Section 3, where the domain is assumed to be one-dimensional ($d=1$). Note that u_0 is determined by the initial value problem

$$\begin{aligned}&\partial_t u_0 + f'(x)\partial_x u_0(x,t) = 0,\\ &u_0(x,0) = \varphi(x).\end{aligned} \tag{D.1}$$

The equation of the characteristic lines is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f'(x(t)) \tag{D.2}$$

which gives

$$t = \int_0^t \frac{\mathrm{d}x(t)}{f'(x(t))} = \int_{x_0(x,t)}^{x} \frac{\mathrm{d}\xi}{f'(\xi)} \tag{D.3}$$

where $x_0 = x_0(x,t)$ is the intercept of the characteristic line passing through the point $(x,t) \in \mathbb{R} \times \mathbb{R}_{\geq 0}$. Therefore,

$$u_0(x,t) = \varphi(x_0(x,t)). \tag{D.4}$$

Using implicit differentiation rules, one can easily deduce from (D.3) that

$$\partial_t x_0(x,t) = -f'(x_0(x,t)), \qquad \partial_x x_0(x,t) = \frac{f'(x_0(x,t))}{f'(x)}$$

with which one easily verifies that (D.4) is the solution of the initial value problem (D.1).

Furthermore, u_1 is determined by the initial value problem

$$\begin{aligned} &\partial_t u_1 + f'(x)\,\partial_x u_1(x,t) = \mathcal{L}_2 u_0(x,t), \\ &u_1(x,0) = 0. \end{aligned} \tag{D.5}$$

Without loss of generality, we will assume $\Sigma = \frac{1}{4}$, which is the variance of a Bernoulli random variable with parameter $p = 1/2$. Using (D.4) and (1.8), we have

$$\begin{aligned} \mathcal{L}_2 u_0(x,t) &= -\frac{1}{2} f'(x) f''(x) \partial_x u_0(x,t) + \frac{1}{8} \partial_x^2 u_0(x,t) \\ &= -\frac{1}{2} f'(x_0(x,t)) \varphi'(x_0(x,t)) f''(x) + \frac{1}{8} \frac{\partial}{\partial x} \left[\frac{f'(x_0(x,t))}{f'(x)} \right] \varphi'(x_0(x,t)) \\ &\quad + \frac{1}{8} \left[\frac{f'(x_0(x,t))}{f'(x)} \right]^2 \varphi''(x_0(x,t)) \end{aligned}$$

in which the middle term in the right-hand side can be further expanded into

$$\begin{aligned} &\frac{1}{8} \varphi'(x_0(x,t)) \frac{f'(x) f''(x_0(x,t)) \partial_x x_0(x,t) - f'(x_0(x,t)) f''(x)}{[f'(x)]^2} \\ &= \frac{1}{8} \varphi'(x_0(x,t)) \frac{f'(x_0(x,t)) [f''(x_0(x,t)) - f''(x)]}{[f'(x)]^2}. \end{aligned}$$

The equation of characteristic lines for (D.5) is the same as (D.3). Using the boundary condition $u_1(x,0) = 0$, we have

$$\begin{aligned} u_1(x(t),t) &= \int_0^t \mathcal{L}_2 u_0(x(t),t)\,\mathrm{d}t \\ &= -\frac{1}{2} f'(x_0) \varphi'(x_0) \int_0^t f''(x(t))\,\mathrm{d}t + \frac{1}{8} f'(x_0) \varphi'(x_0) \int_0^t \frac{f''(x_0) - f''(x(t))}{[f'(x(t))]^2}\,\mathrm{d}t \\ &\quad + \frac{1}{8} [f'(x_0)]^2 \varphi''(x_0) \int_0^t \frac{\mathrm{d}t}{[f'(x(t))]^2} \\ &=: (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}), \end{aligned}$$

where we adopted the simplifying notation $x_0 \equiv x_0(x(t),t)$ for the constant along the characteristic line $x = x(t)$. By means of (D.2), we can further simplify the three terms on the right-hand side:

$$\begin{aligned}
\text{(I)} &= -\frac{1}{2} f'(x_0)\varphi'(x_0) \int_{x_0}^{x} \frac{\mathrm{d}}{\mathrm{d}\xi} \log f'(\xi)\, \mathrm{d}\xi = -\frac{1}{2} f'(x_0)\varphi'(x_0) \log \frac{f'(x)}{f'(x_0)}, \\
\text{(II)} &= \frac{1}{8} f'(x_0)\varphi'(x_0) \int_{x_0}^{x} \frac{f''(x_0) - f''(\xi)}{[f'(\xi)]^3}\, \mathrm{d}\xi, \\
\text{(III)} &= \frac{1}{8} [f'(x_0)]^2 \varphi''(x_0) \int_{x_0}^{x} \frac{\mathrm{d}\xi}{[f'(\xi)]^3}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
u_1(x,t) = &-\frac{1}{2} f'(x_0(x,t))\varphi'(x_0(x,t)) \log \frac{f'(x)}{f'(x_0(x,t))} \\
&+ \frac{1}{8} f'(x_0(x,t))\varphi'(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{f''(x_0(x,t)) - f''(\xi)}{[f'(\xi)]^3}\, \mathrm{d}\xi \\
&+ \frac{1}{8} [f'(x_0(x,t))]^2 \varphi''(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{\mathrm{d}\xi}{[f'(\xi)]^3}.
\end{aligned} \tag{D.6}$$

Alternatively, we can also write u_1 in the following equivalent form:

$$\begin{aligned}
u_1(x,t) = &-\frac{1}{2} f'(x_0(x,t))\varphi'(x_0(x,t)) \log \frac{f'(x)}{f'(x_0(x,t))} \\
&+ \frac{1}{8} f'(x_0(x,t)) f''(x_0(x,t))\varphi'(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{\mathrm{d}\xi}{[f'(\xi)]^3} \\
&- \frac{1}{8} f'(x_0(x,t))\varphi'(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{f''(\xi)}{[f'(\xi)]^3} \mathrm{d}\xi \\
&+ \frac{1}{8} [f'(x_0(x,t))]^2 \varphi''(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{\mathrm{d}\xi}{[f'(\xi)]^3} \\
= &-\frac{1}{2} f'(x_0(x,t))\varphi'(x_0(x,t)) \log \frac{f'(x)}{f'(x_0(x,t))} \\
&+ \frac{1}{8} f'(x_0(x,t)) f''(x_0(x,t))\varphi'(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{\mathrm{d}\xi}{[f'(\xi)]^3} \\
&- \frac{1}{16} f'(x_0(x,t))\varphi'(x_0(x,t)) \left\{ \frac{1}{[f'(x_0(x,t))]^2} - \frac{1}{[f'(x)]^2} \right\} \\
&+ \frac{1}{8} [f'(x_0(x,t))]^2 \varphi''(x_0(x,t)) \int_{x_0(x,t)}^{x} \frac{\mathrm{d}\xi}{[f'(\xi)]^3}.
\end{aligned} \tag{D.7}$$

REFERENCES

[1] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat., 22(3):400–407, 1951. 1

[2] T. Zhang, *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, Proceedings of the 21st International Conference on Machine Learning, 116, 2004. 1

[3] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, *Stochastic convex optimization*, The 22nd Conference on Learning Theory (COLT), June 18–21, 2009. 1

[4] E. Moulines and F.R. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Proceedings of the 24th International Conference on Neural Information Processing Systems, 451–459, 2011. 1

[5] O. Shamir and T. Zhang, *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes*, in International Conference on Machine Learning, 71–79, 2013. 1

[6] F. Bach and E. Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$*, Adv. Neural Inf. Process. Syst. 26, 773–781, 2013. 1

[7] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, and K. Macherey, *Google's neural machine translation system: Bridging the gap between human and machine translation*, arXiv preprint, arXiv:1609.08144, 2016. 1

[8] N.S. Keskar and R. Socher, *Improving generalization performance by switching from Adam to SGD*, arXiv preprint, arXiv:1712.07628, 2017. 1

[9] A.C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, *The marginal value of adaptive gradient methods in machine learning*, in I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Adv. Neural Inf. Process. Syst. 30, Curran Associates, Inc., 30:4148–4158, 2017. 1

[10] B.T. Polyak and A.B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30(4):838–855, 1992. 1

[11] A. Rakhlin, O. Shamir, and K. Sridharan, *Making gradient descent optimal for strongly convex stochastic optimization*, in International Conference on Machine Learning, 12:1571–1578, 2012. 1

[12] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Proceedings of the 26th International Conference on Neural Information Processing Systems, 315–323, 2013. 1

[13] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, Proceedings of the 27th International Conference on Neural Information Processing Systems, 1646–1654, 2014. 1

[14] S.J Reddi, A. Hefny, S. Sra, B. Poczos, and A.J. Smola, *On variance reduction in stochastic gradient descent and its asynchronous variants*, Proceedings of the 28th International Conference on Neural Information Processing Systems, 2647–2655, 2015. 1

[15] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo, *A globally convergent incremental newton method*, Math. Program., 151(1):283–313, 2015. 1

[16] A. Dieuleveut, A. Durmus, and F. Bach, *Bridging the gap between constant step size stochastic gradient descent and Markov chains*, accepted by Ann. Stat., arXiv preprint, arXiv:1707.06386, 2017. 1, 1, 1.1, 2, 2.1

[17] A. Jentzen, B. Kuckuck, A. Neufeld, and P. von Wurstemberger, *Strong error analysis for stochastic gradient descent optimization algorithms*, arXiv preprint, arXiv:1801.09324, 2018. 1, 1

[18] Q. Li, C. Tai, and W. E., *Stochastic modified equations and adaptive stochastic gradient algorithms*, Proceedings of the 34th International Conference on Machine Learning, 70:2101-2110, 2017. 1, 1, 1.1

[19] Q. Li, C. Tai, and W.E., *Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations*, J. Mach. Learn. Res., 20(40):1–47, 2019. 1, 1

[20] Y. Feng, L. Li, and J.-G. Liu, *Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations*, Commun. Math. Sci., 16(3):777–789, 2018. 1, 1, 1.1, 1.1, 2.1, 2.1, 2.3

[21] W. Hu, C.J. Li, L. Li, and J.-G. Liu, *On the diffusion approximation of nonconvex stochastic gradient descent*, Ann. Math. Sci. Appl., 4(1):3–32, 2019. 1, 1.1

[22] A. Debussche and E. Faou, *Weak backward error analysis for SDEs*, SIAM J. Numer. Anal., 50(3):1735–1752, 2012. 1, 1.1

[23] T. Shardlow, *Modified equations for stochastic differential equations*, BIT Numer. Math., 46(1):111–125, 2006. 1, 1.1

[24] A. Abdulle, D. Cohen, G. Vilmart, and K.C. Zygalakis, *High weak order methods for stochastic differential equations based on modified equations*, SIAM J. Sci. Comput., 34(3):A1800–A1823, 2012. 1, 1.1, 1.1

[25] A. Abdulle, G. Vilmart, and K.C. Zygalakis, *High order numerical approximation of the invariant measure of ergodic SDEs*, SIAM J. Numer. Anal., 52(4):1600–1622, 2014. 1, 1.1, 1.1

[26] M. Kopec, *Weak backward error analysis for overdamped Langevin processes*, IMA J. Numer. Anal., 35(2):583–614, 2014. 1, 1.1

[27] M. Kopec, *Weak backward error analysis for Langevin process*, BIT Numer. Math., 55(4):1057–1103, 2015. 1, 1.1

[28] J. Sirignano and K. Spiliopoulos, *Stochastic gradient descent in continuous time*, SIAM J. Finan-

cial Math., 8(1):933–961, 2017. 1
[29] J. Sirignano and K. Spiliopoulos, *Stochastic gradient descent in continuous time: A central limit theorem*, arXiv preprint, arXiv:1710.04273, 2017. 1
[30] S. Fiori, *Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial*, J. Mach. Learn. Res., 6:743–781, 2005. 1.1
[31] U. Helmke and J.B. Moore, *Optimization and Dynamical Systems*, Springer Science & Business Media, 2012. 1.1
[32] H.-B. Dörr, E. Saka, and C. Ebenbauer, *A smooth vector field for quadratic programming*, Proc. IEEE Conf. Decis. Control, 2515–2520, 2012. 1.1
[33] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin, *Sparse recovery via differential inclusions*, Appl. Comput. Harmonic Anal., 41(2):436–469, 2016. 1.1
[34] W. Su, S. Boyd, and E.J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17(153):1–43, 2016. 1.1
[35] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer, Berlin, Heidelberg, Sixth Edition, 2003. 1.1
[36] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, *Entropy-SGD: Biasing gradient descent into wide valleys*, J. Stat. Mech., 2019(12):124018, 2019. 1.1
[37] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier, *Deep relaxation: partial differential equations for optimizing deep neural networks*, Res. Math. Sci., 5(3):30, 2018. 1.1
[38] P. Jain, S.M. Kakade, R. Kidambi, P. Netrapalli, V.K. Pillutla, and A. Sidford, *A Markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares)*, in S. Lokam and R. Ramanujam (eds.), 37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, Leibniz International Proceedings in Informatics (LIPIcs), 93:2:1–2:10, 2018. 1.1
[39] S. Mandt, M.D. Hoffman, and D.M. Blei, *Stochastic gradient descent as approximate Bayesian inference*, J. Mach. Learn. Res., 18(134):1–35, 2017. 1.1
[40] S. Liang, R. Sun, J.D. Lee, and R. Srikant, *Adding one neuron can eliminate all bad local minima*, in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 31:4355–4365, 2018. 2
[41] P. Diaconis and D. Freedman, *Iterated random functions*, SIAM Rev., 41(1):45–76, 1999. 3.1
[42] C. Villani, *Topics in Optimal Transportation*, Graduate Studies in Mathematics, Amer. Math. Soc., Providence, RI, 58, 2003. B