Stochastic Modified Equations: A Unified Framework for Analyzing Gradient Descent and Its Variants

Chris Junchi Li[◇]

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

October 8, 2024

Abstract

Stochastic gradient algorithms are widely used in optimization tasks for machine learning and other domains, particularly when dealing with large datasets or non-convex objectives. This paper introduces a systematic framework based on stochastic modified equations (SME) to analyze the precise dynamics of stochastic gradient algorithms (SGAs). The SME approach approximates discrete-time stochastic algorithms using continuous-time stochastic differential equations, enabling the study of their weak approximations and long-term behavior. We rigorously derive order-1 and order-2 weak approximations for popular gradient-based algorithms such as Stochastic Gradient Descent (SGD) and Momentum-based SGD (MSGD), highlighting the impact of variance-induced noise on the learning dynamics. This unified approach relaxes regularity assumptions and applies to more general stochastic processes. Numerical experiments demonstrate the accuracy of the SME-based analysis in predicting the long-term performance and stability of these optimization algorithms.

Keywords: Stochastic Gradient Algorithms, Stochastic Modified Equations, Weak Approximations, Variance-Induced Noise, Non-Convex Optimization

1 Introduction

Stochastic gradient algorithms (SGAs) are essential tools in machine learning and optimization, particularly in large-scale and non-convex problems. These algorithms provide an efficient means of approximating solutions to problems where evaluating the full gradient is computationally prohibitive. While the stochastic nature of these algorithms introduces noise, which can lead to variance-induced divergence, it also provides the flexibility to escape poor local minima and find better solutions in practice.

However, despite their popularity, a systematic understanding of the precise dynamics of SGAs has been lacking. Discrete-time algorithms such as Stochastic Gradient Descent (SGD) and its variants are often analyzed using disparate techniques, leading to fragmented results across the literature. A continuous-time approximation via stochastic differential equations (SDEs) has the potential to unify the analysis of these algorithms, offering a deeper understanding of their dynamical properties.

In this paper, we aim to bridge this gap by employing the framework of stochastic modified equations (SME) to study SGAs. The SME framework allows us to approximate the discrete-time dynamics of SGAs with continuous-time SDEs, providing a more systematic approach to studying their long-term behavior, convergence rates, and stability.

Mathematically, SGAs are often used to solve optimization problems of the form

$$\min_{x\in\mathbb{R}^d} \quad f(x) := \mathbb{E} f_\gamma(x) \tag{1}$$

where $\{f_r : r \in \Gamma\}$ is a family of functions from $\mathbb{R}^d$ to $\mathbb{R}$ and γ is a Γ-valued random variable, with respect to which the expectation is taken (these notions will be made precise in the following sections). For empirical loss minimization in supervised learning applications, γ is usually a uniform random variable taking values in $\Gamma = \{1, 2, \ldots, n\}$. In this case, f is the total empirical loss function and f_r, $r \in \Gamma$ are the loss function due to the r^{th} training sample. In this paper, we shall consider the general situation of a expectation over arbitrary index sets and distributions.

Solving (1) using the standard gradient descent (GD) on x gives the iteration scheme

$$x_{k+1} = x_k - \eta \nabla \mathbb{E} f_\gamma(x_k), \tag{2}$$

for $k \geq 0$ and η is a small positive step-size known as the learning rate. Note that this requires the evaluation of the gradient of an expectation, which can be costly (in this empirical risk minimization case, this happens when n is large). In its simplest form, the stochastic gradient descent (SGD) algorithm replaces the expectation of the gradient with a sampled gradient, i.e.

$$x_{k+1} = x_k - \eta \nabla f_{\gamma_k}(x_k) \tag{3}$$

where each γ_k is an independent and identically distributed (i.i.d.) random variable with the same distribution as γ. Under mild conditions, we then have $\mathbb{E}[\nabla f_{\gamma_k}(x_k)|x_k] = \nabla \mathbb{E} f(x_k)$. In other words, (3) is a sampled version of (2).

In the literature, many convergence results are available for SGD and its variants [Shamir and Zhang(2013), Moulines and Bach(2011), Needell et al.(2014), Xiao and Zhang(2014), Shalev-Shwartz and Zhang(2014), Bach and Moulines(2013), Défossez and Bach(2015)]. However, it is often the case that different analysis techniques must be adopted for different variants of the algorithms and there generally lacked a systematic approach to study their precise dynamical properties. In [Li et al.(2015)], a general approach was introduced to address this problem, in which discrete-time stochastic gradient algorithms are approximated by continuous-time stochastic differential equations with the noise term depending on a small parameter (the learning rate). This can be viewed as a generalization of the method of modified equations [Hirt(1968), Noh and Protter(1960), Daly(1963), Warming and Hyett(1974)] to the stochastic setting, and allows one to employ tools from stochastic calculus to systematically analyze the dynamics of stochastic gradient algorithms. The stochastic modified equations (SME) approach was further developed in [Li et al.(2017)], where a weak approximation result for the SGD was proved in a finite-sum-objective setting.

The present series of papers builds on the earlier work of [Li et al.(2015), Li et al.(2017)] and aims to establish the framework of stochastic modified equations and their applications in greater generality and depth, and highlight the advantages of this systematic framework for studying stochastic gradient algorithms using continuous-time methods. As the first in the series, this paper will focus on mathematical aspects, namely the main approximation theorems relating stochastic gradient algorithms to stochastic modified equations in the form of weak approximations. These generalize the approximation results in [Li et al.(2017)] in various aspects. In a subsequent paper in the series, we will discuss the application of this formalism to adaptive stochastic gradient algorithms and related problems.

Contributions This paper introduces the stochastic modified equations (SME) framework for analyzing stochastic gradient algorithms (SGAs). We rigorously derive weak approximations of order-1 and order-2 for key algorithms like Stochastic Gradient Descent (SGD) and its momentum variants. By leveraging the SME approach, we uncover the dynamics of SGAs and the impact of variance-induced noise on their stability and long-term behavior. Numerical experiments confirm

the accuracy of this framework in predicting the performance of these algorithms under various conditions.

Organization of the Paper Section 2 reviews related work, focusing on continuous-time approximations of stochastic gradient algorithms. In Section 3, we introduce the SME framework and its mathematical foundations. Section 4 presents our main results, proving key weak approximations that relate discrete stochastic algorithms to continuous stochastic processes, allowing us to derive SMEs for gradient-based algorithms. Section 5 applies the SME framework to analyze the dynamics of SGAs in a practical optimization task, validated through numerical experiments. Section 6 concludes with discussions on future directions. Detailed proofs are provided in the appendix, assuming familiarity with stochastic calculus and probability theory. For background, readers may consult texts such as [Durrett(2010)] and [Oksendal(2013)].

1.1 Notation

In this paper, we adhere wherever possible to the following notation. Dimensional indices are written as subscripts with a bracket to avoid confusion with other sequential indices (e.g. time, iteration number), which do not have brackets. When more than one indices are present, we separate them with a comma, e.g. $x_{k,(i)}$ is the i-th coordinate of the vector x_k, the k^{th} member of a sequence. We adopt the Einstein's summation convention, where repeated (spatial) indices are summed, i.e. $x_{(i)}x_{(i)} := \sum_{i=1}^{d} x_{(i)}x_{(i)}$. For a matrix A, we denote by $\lambda(A) = \{\lambda_1(A), \lambda_2(A), \dots\}$ the set of eigenvalues of A. If A is Hermitian, then the eigenvalues are ordered so that $\lambda_1(A)$ denotes a maximum eigenvalue. We denote the usual Euclidean norm by $|\cdot|$ and for higher rank tensors, we use the same notation to denote the flattened vector norms (e.g. for matrices it will be the Frobenius norm). The $\wedge$ symbols denotes the minimum operator, i.e. $a \wedge b := \min(a, b)$.

For a probability space (or generally, a measure space) $(\Omega, \mathcal{F}, \mathbb{P})$, the symbol $\mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$, $p \in (1, \infty)$ denotes the usual Lebesgue spaces, i.e. $u \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ if

$$\|u\|^p_{\mathcal{L}^p(\Omega,\mathcal{F},\mathbb{P})} := \int_\Omega |u(\omega)|^p d\mathbb{P}(\omega) \equiv \mathbb{E}|u|^p < \infty$$

When the underlying probability space is obvious, we use the shorthand $\mathcal{L}^p(\Omega) \equiv \mathcal{L}(\Omega, \mathcal{F}, \mathbb{P})$. In addition, when $\Omega = \mathbb{R}^d$, we also write the local $\mathcal{L}^p$ spaces as $\mathcal{L}^p_{\text{loc}}(\mathbb{R}^d)$, which contains u for which $|u|^p$ is integrable on compact subsets of $\mathbb{R}^d$.

Finally, we note that in the proofs of various results, we typically use the letter C (whose value may change across results) to denote a generic positive constant. This is usually independent of the learning rate η, but if not explicitly stated otherwise, it may depend on e.g. Lipschitz constants, ambient dimensions, etc.

2 Related work

In this section, we discuss several related works on analyzing discrete-time algorithms using continuous-time approaches. The idea of approximating discrete-time stochastic algorithms by continuous equations dates back to the large body of work known as stochastic approximation theory [Kushner and Yin(2003), Ljung et al.(2012)]. These typically establish law of large numbers type results where the limiting equation is an ODE, which can then be used to prove powerful convergence results for the stochastic algorithms under consideration. A notion of convergence in distribution, similar to a

central limit theorem, was also studied for the purpose of estimating the rate of convergence of the ODE methods [Kushner(1978), Kushner and Shwartz(1984), Kushner and Clark(2012)], where connections between leading order perturbations and Ornstein-Uhlenbeck (OU) processes are established. However, these estimates are not systematically used to systematically study the dynamics of stochastic gradient algorithms.

As far as the authors are aware, the first work on using stochastic differential equations to study the precise properties of stochastic gradient algorithms are the independent works of [Li et al.(2015)] and [Mandt et al.(2015)]. In [Li et al.(2015)], a systematic framework of SDE approximation of SGD and SGD with momentum are derived and applied to study dynamical properties of the stochastic algorithms as well as adaptive parameter tuning schemes. These go beyond OU process approximations and this distinction is important since the OU process is not always the appropriate stochastic approximation in general settings (See Sec. 4.2 of this paper). In [Mandt et al.(2015)], a similar procedure is employed to derive a SDE approximation for the SGD, from which issues such as choice of learning rates are studied. Although the concrete analysis in [Mandt et al.(2015)] is on the restricted case of constant diffusion matrices leading to OU processes, the essential ideas on the general leading order approximation are also discussed.

It is important to note that the approximation arguments in both [Li et al.(2015)] and [Mandt et al.(2015)] are heuristic from a mathematical point of view. In [Li et al.(2017)], the SME approximation is rigorously proved in the finite-sum-objective case with strong regularity conditions, and further asymptotic analysis and tuning algorithms are studied. The SME approach has subsequently been utilized to study variants of stochastic gradient algorithms, including those in the distributed optimization setting [An et al.(2018)]. The work of [Mandt et al.(2015)] is further developed in [Mandt et al.(2016), Mandt et al.(2017)], with applications such as the development scalable MCMC algorithms.

The present paper builds on the earlier work of [Li et al.(2015), Li et al.(2017)], but focuses on extending and solidifying the mathematical aspects. In particular, we present an entirely rigorous and self-contained mathematical formulation of the SME framework that applies to more general algorithms (including momentum SGD and stochastic Nesterov's accelerated gradient method) and more general objectives (expectation over random functions, instead of just a finite-sum). Moreover, various regularity conditions in [Li et al.(2017)] have been relaxed. The main approximation procedure is inspired by the seminal works of [Milstein(1986), Milstein(1975)] in numerical analysis of stochastic differential equations, but lower regularity conditions are required in our case due to the presence of the small noise parameter, which allows for better truncation of Itô-Taylor expansions. The mathematical analysis of the SME-type approximation for the SGD was also performed in [Feng et al.(2017), Hu et al.(2017)] using semi-group approaches, although the smoothness requirements presented there are greater than those established using the current methods. Lastly, the Nesterov's accelerated gradient SME we derive in Sec. 4.4 can be viewed as a generalization of the ODE approach in [Su et al.(2014)] to stochastic gradients, and we show that the presence of noise gives additional features to the dynamics. Finally, we note that continuous-time approximations that establish links between optimization, calculus of variations and symplectic integration has been studied in [Wibisono et al.(2016), Betancourt et al.(2018)].

3 Stochastic modified equations

We now introduce the stochastic modified equations framework. The starting motivation is the observation that GD iterations is a (Euler) discretization of the continuous-time, ordinary differential

equation

$$\frac{dx}{dt} = -\nabla f(x) \tag{4}$$

and studying (4) can give us important insights to the dynamics of the discrete-time algorithm for small enough learning rates. The natural question when extending this to SGD is, ***what is the right continuous-time equation to consider?*** Below, we begin with some heuristic considerations.

3.1 Heuristic motivations

we rewrite the SGD iteration (3) as

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{\eta} V_k(x_k, \gamma_k) \tag{5}$$

where $V_k(x_k, \gamma_k) = \sqrt{\eta}(\nabla f(x_k) - \nabla f_{\gamma_k}(x_k))$ is a d-dimensional random vector. A straightforward calculation shows that

$$\begin{aligned} &\mathbb{E}[V_k|x_k] = 0 \\ &\mathrm{cov}[V_k, V_k|x_k] = \eta\Sigma(x_k), \\ &\Sigma(x_k) := \mathbb{E}[(\nabla f_{\gamma_k}(x_k) - \nabla f(x_k))(f_{\gamma_k}(x_k) - \nabla f(x_k))^T|x_k] \end{aligned} \tag{6}$$

i.e. conditional on x_k, $V_k(x_k)$ has 0 mean and covariance $\eta\Sigma(x_k)$. Here, Σ is simply the conditional covariance of the stochastic gradient approximation ∇f_γ of ∇f.

Now, consider a time-homogeneous Itô stochastic differential equation (SDE) of the form

$$dX_t = b(X_t)dt + \sqrt{\eta}\sigma(X_t)dW_t \tag{7}$$

where $X_t \in \mathbb{R}^d$ for $t \geq 0$ and W_t is a standard d-dimensional Wiener process. The function $b : \mathbb{R}^d \to \mathbb{R}^d$ is known as the drift and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the diffusion matrix. The key observation is that if we apply the Euler discretization with step-size η to (7), approximating $X_{k\eta}$ by $\widehat{X}_k$, we obtain the following discrete iteration for the latter:

$$\widehat{X}_{k+1} = \widehat{X}_k + \eta b(\widehat{X}_k) + \eta\sigma(\widehat{X}_k)Z_k \tag{8}$$

where $Z_k := W_{(k+1)\eta} - W_{k\eta}$ are d-dimensional i.i.d. standard normal random variables. Comparing with (5), if we set $b = -\nabla f$, $\sigma(x) = \Sigma(x)^{1/2}$ and identify t with $k\eta$, we then have matching first and second conditional moments. Hence, this motivates the approximating equation

$$dX_t = -\nabla f(X_t)dt + (\eta\Sigma(X_t))^{1/2}dW_t \tag{9}$$

Note that as this heuristic argument shows, the presence of the small parameter $\sqrt{\eta}$ on the diffusion term is necessary to model the fact that when learning rate decreases, the fluctuations to the SGA iterates must also decrease.

The immediate mathematical question is then: ***in what sense is an SDE like*** (9) ***an approximation of*** (3)***?*** Let us now establish the precise mathematical framework in which we can answer this question.

3.2 The mathematical framework

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a sufficiently rich probability space and $(\Gamma, \mathcal{F}_\Gamma)$ be a measure space representing the index space for our random objectives. Let $\gamma : \Omega \to \Gamma$ be a random variable and $(r, x) \mapsto f_r(x)$ a measurable mapping from $\Gamma \times \mathbb{R}^d$ to $\mathbb{R}$. Hence, for each x, $f_\gamma(x)$ is a random variable. Throughout this paper, we assume the follow facts about $f_\gamma(x)$:

Assumption 1. *The random variable $f_\gamma(x)$ satisfies*

(i) $f_\gamma(x) \in \mathcal{L}^1(\Omega)$ *for all* $x \in \mathbb{R}^d$

(ii) $f_\gamma(x)$ *is continuously differentiable in* x *almost surely and for each* $R > 0$*, there exists a random variable* $M_{R,\gamma}$ *such that* $\max_{|x|\le R} |\nabla f_\gamma(x)| \le M_{R,\gamma}$ *almost surely, with* $\mathbb{E}|M_{R,\gamma}| < \infty$

(iii) $\nabla f_\gamma(x) \in \mathcal{L}^2(\Omega)$ *for all* $x \in \mathbb{R}^d$

Note that in the empirical risk minimization case where Γ is finite, the conditions above are often trivially satisfied. Condition (i) in Assumption 1 allows us to define the total objective function we would like to minimize as the expectation

$$f(x) := \mathbb{E} f_\gamma(x) \equiv \int_\Omega f_{\gamma(\omega)}(x) d\mathbb{P}(\omega) \tag{10}$$

Moreover, Assumption 1 (ii) implies via the dominated convergence theorem that $\mathbb{E}\nabla f_\gamma = \nabla \mathbb{E} f_\gamma \equiv \nabla f$. Now, let $\{\gamma_k : k = 0, 1, \dots\}$ be a sequence of i.i.d. Γ-valued random variables with the same distribution as γ. Let $x_0 \in \mathbb{R}^d$ be fixed and define the generalized stochastic gradient iteration as the stochastic process

$$x_{k+1} = x_k + \eta h(x_k, \gamma_k, \eta) \tag{11}$$

for $k \ge 0$, where $h : \mathbb{R}^d \times \Gamma \times \mathbb{R} \to \mathbb{R}^d$ is a measurable function and $\eta > 0$ is the learning rate. In the simple case of SGD, we have $h(x, r, \eta) = -\nabla f_r(x)$, but we shall consider the generalized version above so that modified equations for SGD variants can also be derived from our approximation theorems.

Next, let us define the class of approximating continuous stochastic processes, which we call stochastic modified equations. Consider the time-homogeneous Itô diffusion process $\{X_t : t \ge 0\}$ represented by the following stochastic differential equation (SDE)

$$dX_t = b(X_t, \eta)dt + \sqrt{\eta}\sigma(X_t, \eta)dW_t, \qquad X_0 = x_0 \tag{12}$$

where $\{W_t : t \ge 0\}$ is a standard d-dimensional Wiener process independent of $\{\gamma_k\}$, $b : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is the approximating drift vector and $\sigma : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d\times d}$ is the approximating diffusion matrix. In the following, we will need to pick b, σ appropriately so that (11) is approximated by (12), the sense of which we now describe.

First, notice that the stochastic process $\{x_k\}$ induces a probability measure on the product space $\mathbb{R}^d \times \mathbb{R}^d \times \cdots$, whereas $\{X_t\}$ induces a probability measure on $\mathcal{C}^0([0, \infty), \mathbb{R}^d)$. Hence, we can only compare their values by sampling a discrete number of points from the latter. Second, the process $\{x_k\}$ is adapted to the filtration generated by $\{\gamma_k\}$ (e.g. in the case of SGD, this is the random sampling of functions in $\{f_r\}$), whereas the process $\{X_t\}$ is adapted to an independent, Wiener filtration. Hence, it is not appropriate to compare individual sample paths. Rather, we define below a sense of weak approximations by comparing the distributions of the two processes.

Definition 1. *Let G denote the set of continuous functions $\mathbb{R}^d \to \mathbb{R}$ of at most polynomial growth, i.e. $g \in G$ if there exists positive integers $\kappa_1, \kappa_2 > 0$ such that*

$$|g(x)| \leq \kappa_1(1 + |x|^{2\kappa_2})$$

for all $x \in \mathbb{R}^d$. Moreover, for each integer $\alpha \geq 1$ we denote by G^α the set of α-times continuously differentiable functions $\mathbb{R}^d \to \mathbb{R}$ which, together with its partial derivatives up to and including order α, belong to G. Note that each G^α is a subspace of $\mathcal{C}^\alpha$, the usual space of α-times continuously differentiable functions. Moreover, if g depends on additional parameters, we say $g \in G^\alpha$ if the constants κ_1, κ_2 are independent of these parameters, i.e. $g \in G^\alpha$ uniformly. Finally, the definition generalizes to vector-valued functions coordinate-wise in the co-domain.

Definition 2. *Let $T > 0$, $\eta \in (0, 1 \wedge T)$, and $\alpha \geq 1$ be an integer. Set $N = \lfloor T/\eta \rfloor$. We say that a continuous-time stochastic process $\{X_t : t \in [0, T]\}$ is an order α weak approximation of a discrete stochastic process $\{x_k : k = 0, \ldots, N\}$ if for every $g \in G^{\alpha+1}$, there exists a positive constant C, independent of η, such that*

$$\max_{k=0,\ldots,N} |\mathbb{E}g(x_k) - \mathbb{E}g(X_{k\eta})| \leq C\eta^\alpha \tag{13}$$

Let us discuss briefly the notion of *weak approximation* as introduced above. These are approximations of the distribution of sample paths, instead of the sample paths themselves. This is enforced by requiring that the expectations of the two processes $\{X_t\}$ and $\{x_k\}$ over a sufficiently large class of test functions to be close. In our definition, the test function class $G^{\alpha+1}$ is quite large, and in particular it includes all polynomials. Thus, Eq. (13) implies in particular that all moments of the two processes become close at the rate of η^α, and hence so must their distributions. The notion of weak approximation must be contrasted with that of *strong approximations*, where one would for example require (in the case of *mean-square approximations*)

$$\left[\mathbb{E}|x_k - X_{k\eta}|^2\right]^{1/2} \leq C\eta^\alpha$$

The above forces the actual sample-paths of the two processes to be close, per realization of the random process, which severely limits its application. In fact, one important advantage of weak approximations is that the approximating SDE process X_t can in fact approximate discrete stochastic processes whose step-wise driving noise is not Gaussian, which is exactly what we need to analyze general stochastic gradient iterations.

4 The approximation theorems

We now present the main approximation theorems. The derivation is based on the following two-step process:

1. We establish a connection between one-step approximation and approximation on a finite time interval.

2. We construct a one-step approximation that is of order $\alpha + 1$, and so the approximation on a finite interval is of order α.

4.1 Relating one-step to N-step approximations

Let us consider generally the question of the relationship between one-step approximations and approximations on a finite interval. Let $T > 0$, $\eta \in (0, 1 \wedge T)$ and $N = \lfloor T/\eta \rfloor$ and recall the general SGA iterations

$$x_{k+1} = x_k + \eta h(x_k, \gamma_k, \eta), \quad x_0 \in \mathbb{R}^d, \quad k = 0, \dots, N \tag{14}$$

and the general candidate family of approximating SDEs

$$dX_t^{\eta,\epsilon} = b(X_t^{\eta,\epsilon}, \eta, \epsilon)dt + \sqrt{\eta}\sigma(X_t^{\eta,\epsilon}, \eta, \epsilon)dW_t, \quad X_0 = x_0, \quad t \in [0, T] \tag{15}$$

where $\epsilon \in (0, 1)$ is a mollification parameter, whose role will become apparent later. To reduce notational clutter and improve readability, unless some limiting procedure is considered, we shall not explicit write the dependence of $X_t^{\eta,\epsilon}$ on η, ϵ and simply denote by X_t the solution of the above SDE. Let us also denote for convenience $\widetilde{X}_k := X_{k\eta}$. Further, let $\{X_t^{x,s} : t \geq s\}$ denote the stochastic process obeying the same equation (15), but with the initial condition $X_s^{x,s} = x$. We similarly write $\widetilde{X}_k^{x,l} := X_{k\eta}^{x,l\eta}$ and denote by $\{x_k^{x,l} : k \geq l\}$ the stochastic process satisfying (14) but with $x_l = x$.

Throughout this section, we assume the following conditions:

Assumption 2. *The functions $b : \mathbb{R}^d \times (0, 1\wedge T) \times (0, 1) \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times (0, 1\wedge T) \times (0, 1) \to \mathbb{R}^{d\times d}$ satisfy:*

1. *Uniform linear growth condition*

$$|b(x, \eta, \epsilon)|^2 + |\sigma(x, \eta, \epsilon)|^2 \leq L^2(1 + |x|^2)$$

   for all $x, y \in \mathbb{R}^d$, $\eta \in (0, 1 \wedge T)$, $\epsilon \in (0, 1)$.

2. *Uniform Lipschitz condition*

$$|b(x, \eta, \epsilon) - b(y, \eta, \epsilon)| + |\sigma(x, \eta, \epsilon) - \sigma(y, \eta, \epsilon)| \leq L|x - y|$$

   for all $x, y \in \mathbb{R}^d$, $\eta \in (0, 1 \wedge T)$, $\epsilon \in (0, 1)$.

Note that 2 implies 1 if there is at least one x where the supremum of b, σ over η, ϵ is finite. In particular, these conditions imply via Thm. 5 that there exists a unique solution to Eq. 15.

Now, let us denote the one-step changes

$$\Delta(x) := x_1^{x,0} - x, \qquad \widetilde{\Delta}(x) := \widetilde{X}_1^{x,0} - x \tag{16}$$

We prove the following result which relates one-step approximations with approximations on a finite time interval.

Theorem 1. *Let $T > 0$, $\eta \in (0, 1\wedge T)$, $\epsilon \in (0, 1)$ and $N = \lfloor T/\eta \rfloor$. Let $\alpha \geq 1$ be an integer. Suppose further that the following conditions hold:*

(i) There exists a function $\rho : (0, 1) \to \mathbb{R}_+$ and $K_1 \in G$ independent of η, ϵ such that

$$\left| \mathbb{E} \prod_{j=1}^{s} \Delta_{(i_j)}(x) - \mathbb{E} \prod_{j=1}^{s} \widetilde{\Delta}_{(i_j)}(x) \right| \leq K_1(x)(\eta\rho(\epsilon) + \eta^{\alpha+1})$$

for $s = 1, 2, \ldots, \alpha$ *and*

$$\mathbb{E} \prod_{j=1}^{\alpha+1} \left| \Delta_{(i_j)}(x) \right| \leq K_1(x) \eta^{\alpha+1}$$

where $i_j \in \{1, \ldots, d\}$.

(ii) For each $m \geq 1$, *the* $2m$*-moment of* $x_k^{x,0}$ *is uniformly bounded with respect to* k *and* η, *i.e. there exists a* $K_2 \in G$, *independent of* η, k, *such that*

$$\mathbb{E}|x_k^{x,0}|^{2m} \leq K_2(x)$$

for all $k = 0, \ldots, N \equiv \lfloor T/\eta \rfloor$.

Then, for each $g \in G^{\alpha+1}$, *there exists a constant* $C > 0$, *independent of* η, ϵ, *such that*

$$\max_{k=0,\ldots,N} |\mathbb{E}g(x_k) - \mathbb{E}g(X_{k\eta})| \leq C(\eta^\alpha + \rho(\epsilon))$$

The proof of Thm. 1 requires a number of technical results that we defer to the appendix. Below, we demonstrate the main ingredients of the proof and refer to the appendix where the proofs of the auxiliary results are fully presented.

Proof. In this proof, since there are many conditioning on the initial condition, to prevent nested superscripts we shall introduce the alternative notation $X_t(x, s) \equiv X_t^{x,s}$, and similarly for $\widetilde{X}_k$ and x_k. Fix $g \in G^{\alpha+1}$ and $1 \leq k \leq N$. We have

$$\mathbb{E}g(X_{k\eta}) = \mathbb{E}g(\widetilde{X}_k) = \mathbb{E}g(\widetilde{X}_k(\widetilde{X}_1, 1)) - \mathbb{E}g(\widetilde{X}_k(x_1, 1)) + \mathbb{E}g(\widetilde{X}_k(x_1, 1))$$

If $k > 1$, by noting that $\widetilde{X}_k(x_1, 1) = \widetilde{X}_k(\widetilde{X}_2(x_1, 1), 2)$, we get

$$\mathbb{E}g(\widetilde{X}_k(x_1, 1)) = \mathbb{E}g(\widetilde{X}_k(\widetilde{X}_2(x_1, 1), 2)) - \mathbb{E}g(\widetilde{X}_k(x_2, 2)) + \mathbb{E}g(\widetilde{X}_k(x_2, 2))$$

Continuing this process, we then have

$$\begin{aligned} \mathbb{E}g(\widetilde{X}_k) = &\sum_{l=1}^{k-1} \mathbb{E}g(\widetilde{X}_k(\widetilde{X}_l(x_{l-1}, l-1), l)) - \mathbb{E}g(\widetilde{X}_k(x_l, l)) \\ &+ \mathbb{E}g(\widetilde{X}_k(x_{k-1}, k-1)) \end{aligned}$$

and hence by subtracting $\mathbb{E}g(x_k) \equiv \mathbb{E}g(x_k(x_{k-1}, k-1))$ we get

$$\begin{aligned} \mathbb{E}g(\widetilde{X}_k) - \mathbb{E}g(x_k) = &\sum_{l=1}^{k-1} \mathbb{E}g(\widetilde{X}_k(\widetilde{X}_l(x_{l-1}, l-1), l)) - \mathbb{E}g(\widetilde{X}_k(x_l, l)) \\ &+ \mathbb{E}g(\widetilde{X}_k(x_{k-1}, k-1)) - \mathbb{E}g(x_k(x_{k-1}, k-1)) \end{aligned}$$

and so

$$\mathbb{E}g(\widetilde{X}_k) - \mathbb{E}g(x_k) = \sum_{l=1}^{k-1} \mathbb{E}\mathbb{E}\left[g(\widetilde{X}_k(\widetilde{X}_l(x_{l-1}, l-1), l)) \middle| \widetilde{X}_l(x_{l-1}, l-1) \right] - \mathbb{E}\mathbb{E}\left[g(\widetilde{X}_k(x_l, l)) \middle| x_l \right]$$

$$+ \mathbb{E}g(\widetilde{X}_k(x_{k-1}, k-1)) - \mathbb{E}g(x_k(x_{k-1}, k-1))$$

Now, let $u(x,s) = \mathbb{E}g(X_{k\eta}(x,s))$. Then, we have

$$\begin{aligned}
|\mathbb{E}g(\widetilde{X}_k) - \mathbb{E}g(x_k)| \leq & \sum_{l=1}^{k-1} |\mathbb{E}u(\widetilde{X}_l(x_{l-1}, l-1), l\eta) - \mathbb{E}u(x_l(x_{l-1}, l-1), l\eta)| \\
& + |\mathbb{E}g(\widetilde{X}_k(x_{k-1}, k-1)) - \mathbb{E}g(x_k(x_{k-1}, k-1))| \\
\leq & \sum_{l=1}^{k-1} \mathbb{E}|\mathbb{E}[u(\widetilde{X}_l(x_{l-1}, l-1), l\eta)|x_{l-1}] - \mathbb{E}[u(x_l(x_{l-1}, l-1), l\eta)|x_{l-1}]| \\
& + \mathbb{E}|\mathbb{E}[g(\widetilde{X}_k(x_{k-1}, k-1))|x_{k-1}] - \mathbb{E}[g(x_k(x_{k-1}, k-1))|x_{k-1}]|
\end{aligned}$$

Using Prop. 1, $u(\cdot, s) \in G^{\alpha+1}$ uniformly in s, t, η and ϵ. Thus, by Assumption (i) and Lem. 10,

$$\begin{aligned}
|\mathbb{E}g(x_k) - \mathbb{E}g(\widetilde{X}_k)| \leq & (\eta\rho(\epsilon) + \eta^{\alpha+1}) \left(\sum_{l=1}^{k-1} \mathbb{E}K_{l-1}(x_{l-1}) + \mathbb{E}K_{k-1}(x_{k-1}) \right) \\
\leq & (\eta\rho(\epsilon) + \eta^{\alpha+1}) \sum_{l=0}^{N} \kappa_{l,1}(1 + \mathbb{E}|x_l|^{2\kappa_{l,2}})
\end{aligned}$$

where in the last line we used moment estimates from Thm. 6. Finally, using Assumption (ii) and the fact that $N \leq T/\eta$, we have

$$|\mathbb{E}g(x_k) - \mathbb{E}g(X_{k\eta})| = |\mathbb{E}g(x_k) - \mathbb{E}g(\widetilde{X}_k)| \leq C(\rho(\epsilon) + \eta^{\alpha})$$

□

4.2 SME for stochastic gradient descent

Thm. 1 allows us to prove the main approximation results for the current paper. In particular, in this section we derive a second-order accurate weak approximation for the simple SGD iterations (3), from which a simpler, first-order accurate approximation also follows. As seen in Thm. 1, we need only verify the conditions (i)-(ii) in order to prove the weak approximation result. These conditions mostly involve moment estimates, which we now perform. To simplify presentation, we introduce the following shorthand. Whenever we write

$$\psi(x) = \psi_0(x) + \eta\psi_1(x) + \mathcal{O}(r(\eta, \epsilon))$$

for some remainder term $r(\eta, \epsilon)$, we mean: there exists $K \in G$ independent of η, ϵ such that

$$|\psi(x) - \psi_0(x) - \eta\psi_1(x)| \leq K(x)r(\eta, \epsilon)$$

Now, let us set in (15)

$$\begin{aligned}
b(x, \eta, \epsilon) &= b_0(x, \epsilon) + \eta b_1(x, \epsilon) \\
\sigma(x, \eta, \epsilon) &= \sigma_0(x, \epsilon)
\end{aligned}$$

where b_0, b_1, σ_0 are functions to be determined. We have the following moment estimate.

Lemma 1. *Let $\widetilde{\Delta}(x)$ be defined as in (16). Suppose further that with $b_0, b_1, \sigma_0 \in G^3$. Then we have*

(i) $\mathbb{E}\widetilde{\Delta}_{(i)}(x) = b_0(x,\epsilon)_{(i)}\eta + [\frac{1}{2}b_0(x,\epsilon)_{(j)}\partial_{(j)}b_0(x,\epsilon)_{(i)} + b_1(x,\epsilon)_{(i)}]\eta^2 + \mathcal{O}(\eta^3)$,

(ii) $\mathbb{E}\widetilde{\Delta}_{(i)}(x)\widetilde{\Delta}_{(j)}(x) = [b_0(x,\epsilon)_{(i)}b_0(x,\epsilon)_{(j)} + \sigma_0(x,\epsilon)_{(i,k)}\sigma_0(x,\epsilon)_{(j,k)}]\eta^2 + \mathcal{O}(\eta^3)$,

(iii) $\mathbb{E}\prod_{j=1}^3 |\widetilde{\Delta}_{(i_j)}(x)| = \mathcal{O}(\eta^3)$.

Proof. To obtain (i)-(iii), we simply apply Lem. 11 with $\psi(z) = \prod_{j=1}^s (z_{(i_j)} - x_{(i_j)})$ for $s = 1, 2, 3$ respectively. □

Next, we estimate the moments of the SGA iterations below.

Lemma 2. *Let $\Delta(x)$ be defined as in (16) with the SGD iterations, i.e. $h(x, r, \eta) = -\nabla f_r(x)$. Suppose that for each $x \in \mathbb{R}^d$, $f \in G^1$. Then,*

(i) $\mathbb{E}\Delta_{(i)}(x) = -\partial_{(i)}f(x)\eta$,

(ii) $\mathbb{E}\Delta_{(i)}(x)\Delta_{(j)}(x) = \partial_{(i)}f(x)\partial_{(j)}f(x)\eta^2 + \Sigma(x)_{(i,j)}\eta^2$,

(iii) $\mathbb{E}\prod_{j=1}^3 |\Delta_{(i_j)}(x)| = \mathcal{O}(\eta^3)$,

where $\Sigma(x) := \mathbb{E}(\nabla f_\gamma(x) - \nabla f(x))(\nabla f_\gamma(x) - \nabla f(x))^T$.

Proof. We have $\Delta(x) = -\eta\nabla f_{\gamma_0}(x)$. Taking expectations, the results then follow. □

We now prove the main approximation theorem for the simple SGD. Before presenting the statement and proof, we shall note a few technical issues that prevents the direct application of Thm. 1 with the moment estimates in Lem.1 and 2. The latter suggest ignoring ϵ and setting

$$b_0(x,\epsilon) = -\nabla f(x), \quad b_1(x,\epsilon) = --\frac{1}{4}\nabla|\nabla f(x)|^2, \quad \sigma_0(x,\epsilon) = \Sigma(x)^{\frac{1}{2}}$$

Then, we would see from Lem.1 and 2 that the SGD and the SDE have matching moments up to $\mathcal{O}(\eta^3)$. The first issue with this approach is that even if $\Sigma(x)$ is sufficiently smooth (which may follow from the regularity of ∇f_γ), the smoothness of $\Sigma(x)^{1/2}$ cannot be guaranteed unless $\Sigma(x)$ is positive-definite, which is often too strong an assumption in practice and excludes interesting cases where $\Sigma(x)$ is a singular diffusion matrix. However, the results in Sec. 4.1 require smoothness. Second, we would like to consider functions f_γ that may not have higher strong derivatives required by the Lemmas, beyond those required to define the modified equation itself. To fix both of these issues, we will use a simple mollifying technique. This is the reason for the inclusion of the ϵ parameter in the results in Sec. 4.1.

Definition 3. *Let us denote by $\nu : \mathbb{R}^d \to \mathbb{R}$, $\nu \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ the standard mollifier*

$$\nu(x) := \begin{cases} C\exp(-\frac{1}{1-|x|^2}) & |x| < 1 \\ 0 & |x| \geq 1, \end{cases}$$

where $C := (\int_{\mathbb{R}^d}\nu(y)dy)^{-1}$ is chosen so that the integral of ν is 1. Further, define $\nu^\epsilon(x) = \epsilon^{-d}\nu(x/\epsilon)$. Let $\psi \in \mathcal{L}^1_{loc}(\mathbb{R}^d)$ be locally integrable, then we may define its mollification by

$$\psi^\epsilon(x) := (\nu^\epsilon * \psi)(x) = \int_{\mathbb{R}^d}\nu^\epsilon(x-y)\psi(y)dy = \int_{\mathcal{B}(0,\epsilon)}\nu^\epsilon(y)\psi(x-y)dy$$

where $\mathcal{B}(z,\epsilon)$ is the d-dimensional ball of radius ϵ centered at z. The mollification of vector (or matrix) valued functions are defined element-wise.

The mollifier has very useful properties. In particular, we will use the following well-known facts (see e.g. [Evans(2010)] for proof)

(i) If $\psi \in \mathcal{L}^1_{\text{loc}}(\mathbb{R}^d)$, then $\psi^\epsilon \in \mathcal{C}^\infty(\mathbb{R}^d)$

(ii) $\psi^\epsilon(x) \to \psi(x)$ as $\epsilon \to 0$ for almost every $x \in \mathbb{R}^d$ (with respect to the Lebesgue measure)

(iii) If ψ is continuous, then $\psi^\epsilon(x) \to \psi(x)$ as $\epsilon \to 0$ uniformly on compact subsets of $\mathbb{R}^d$

Next, we make use of the idea of weak derivatives.

Definition 4. *Let $\Psi \in \mathcal{L}^1_{loc}(\mathbb{R}^d)$ and J be a multi-index of order $|J|$. Suppose that there exists a $\psi \in \mathcal{L}^1_{loc}(\mathbb{R}^d)$ such that*

$$\int_{\mathbb{R}^d} \Psi(x)\nabla^J\phi(x)dx = (-1)^{|J|}\int_{\mathbb{R}^d}\psi(x)\phi(x)dx$$

for all $\phi \in \mathcal{C}^\infty_c$. Then, we call ψ the order J weak derivative of Ψ and write $D^J\Psi = \psi$. Note that when it exists, the weak derivative is unique almost everywhere and if Ψ is differentiable, $\nabla^J\Psi = D^J\Psi$ almost everywhere [Evans(2010)].

The introduction of weak derivatives motivates the definition of the weak version of the function spaces G^α.

Definition 5. *For $\alpha \geq 1$, we define the space G^α_w to be the subspace of $\mathcal{L}^1_{loc}(\mathbb{R}^d)$ such that if $g \in G^\alpha_w$, then g has weak derivatives up to order α and for each multi-index J with $|J| \leq \alpha$, there exists positive integers κ_1, κ_2 such that*

$$|D^J g(x)| \leq \kappa_1(1 + |x|^{2\kappa_2}) \text{ for a.e. } x \in \mathbb{R}^d$$

As in Def. 1, if g depends on additional parameters, we say that $g \in G^\alpha_w$ if the above constants do not depend on the additional parameters. Also, vector-valued g are defined as above element-wise in the co-domain. Note that G^α_w is a subspace of the Sobolev space $W^{\alpha,1}_{loc}$.

Theorem 2. *Let, $T > 0$, $\eta \in (0, 1 \wedge T)$ and set $N = \lfloor T/\eta \rfloor$. Let $\{x_k : k \geq 0\}$ be the SGD iterations defined in (3). Suppose the following conditions are met:*

(i) $f \equiv \mathbb{E}f_\gamma$ is twice continuously differentiable, $\nabla|\nabla f|^2$ is Lipschitz, and $f \in G^4_w$.

(ii) ∇f_γ satisfies a Lipschitz condition:

$$|\nabla f_\gamma(x) - \nabla f_\gamma(y)| \leq L_\gamma|x - y| \quad a.s$$

for all $x, y \in \mathbb{R}^d$, where L_γ is a random variable which is positive a.s. and $\mathbb{E}L^m_\gamma < \infty$ for each $m \geq 1$.

Define $\{X_t : t \in [0,T]\}$ as the stochastic process satisfying the SDE

$$dX_t = -\nabla(f(X_t) + \frac{1}{4}\eta|\nabla f(X_t)|^2)dt + \sqrt{\eta}\Sigma(X_t)^{1/2}dW_t \qquad X_0 = x_0 \tag{17}$$

with $\Sigma(x) = \mathbb{E}(\nabla f_\gamma(x) - \nabla f(x))(\nabla f_\gamma(x) - \nabla f(x))^T$. *Then,* $\{X_t : t \in [0,T]\}$ *is an order-2 weak approximation of the SGD, i.e. for each* $g \in G^3$*, there exists a constant* $C > 0$ *independent of* η *such that*

$$\max_{k=0,\dots,N} |\mathbb{E}g(x_k) - \mathbb{E}g(X_{k\eta})| \leq C\eta^2$$

Proof. First, we check that Eq. (17) admits a unique solution, which amounts to checking the conditions in Thm. 5. Note that the Lipschitz condition (ii) implies ∇f is Lipschitz with constant $\mathbb{E}L_\gamma$. To see that $\Sigma(x)^{1/2}$ is also Lipschitz, observe that $u(x) := \nabla f_\gamma(x) - \nabla f(x)$ is Lipschitz (in the sense of (ii), with constant at most $L_\gamma + \mathbb{E}L_\gamma$), and

$$\begin{aligned}|\Sigma(x)^{1/2} - \Sigma(y)^{1/2}| =& \left| \|[u(x)u(x)^T]^{1/2}\|_{\mathcal{L}^2(\Omega)} - \|[u(y)u(y)^T]^{1/2}\|_{\mathcal{L}^2(\Omega)} \right| \\ \leq& \|[u(x)u(x)^T]^{1/2} - [u(y)u(y)^T]^{1/2}\|_{\mathcal{L}^2(\Omega)}\end{aligned}$$

Moreover, observe that for vectors $u \in \mathbb{R}^d$ the mapping $u \mapsto (uu^T)^{1/2} = uu^T/|u|$ is Lipschitz, which implies

$$|\Sigma(x)^{1/2} - \Sigma(y)^{1/2}| \leq L'\|u(x) - u(y)\|_{\mathcal{L}^2(\Omega)} \leq L''|x - y|$$

The Lipschitz conditions on the drift and the diffusion matrix imply uniform linear growth, so by Thm. 5, Eq. (17) admits a unique solution.

For each $\epsilon \in (0,1)$, define the mollified functions

$$b_0(x,\epsilon) = -\nu^\epsilon * \nabla f(x), \quad b_1(x,\epsilon) = -\frac{1}{4}\nu^\epsilon * (\nabla|\nabla f(x)|^2), \quad \sigma_0(x,\epsilon) = \nu^\epsilon * \Sigma(x)^{1/2}$$

Observe that $b_0 + \eta b_1, \sigma_0$ satisfies a Lipschitz condition in x uniformly in η, ϵ. To see this, note that for any Lipschitz function ψ with constant L, we have

$$|\nu^\epsilon * \psi(x) - \nu^\epsilon * \psi(y)| \leq \int_{\mathcal{B}(0,\epsilon)} \nu^\epsilon(z)|\psi(x-z) - \psi(y-z)|dz \leq L|x-y|$$

which proves $b_0 + \eta b_1$ and σ_0 are uniformly Lipschitz. Similarly, the linear growth condition follows. Hence, we may define a family of stochastic processes $\{X_t^\epsilon : \epsilon \in (0,1)\}$ satisfying

$$dX_t^\epsilon = b_0(X_t^\epsilon, \epsilon) + \eta b_1(X_t^\epsilon, \epsilon) + \sqrt{\eta}\sigma_0(X_t^\epsilon, \epsilon)dW_t \qquad X_0^\epsilon = x_0$$

which each admits a unique solution by Thm. 5. Now, we claim that $b_0(\cdot,\epsilon), b_1(\cdot,\epsilon), \sigma_0(\cdot,\epsilon) \in G^3$ uniformly in ϵ. To see this, simply observe that mollifications are smooth, and moreover, the polynomial growth is satisfied since $\nu^\epsilon * D^J\psi = \nabla^J(\nu^\epsilon * \psi)$ and furthermore, if $\psi \in G$, then we have

$$\begin{aligned}|\psi^\epsilon(x)| \leq& \int_{\mathcal{B}(0,\epsilon)} \nu^\epsilon(y)|\psi(x-y)|dy \\ \leq& \kappa_1 \left(1 + 2^{2\kappa_2 - 1}|x|^{2\kappa_2} + 2^{2\kappa_2-1}\frac{1}{\epsilon^d}\int_{\mathcal{B}(0,\epsilon)} |y|^{2\kappa_2}dy\right)\end{aligned}$$

But $\int_{\mathcal{B}(0,\epsilon)} |y|^{2\kappa_2}dy \leq \text{Vol}(\mathcal{B}(0,\epsilon)) = C\epsilon^d$, where C is independent of ϵ. This shows that $\psi^\epsilon \in G$ uniformly in ϵ. This immediately implies that $b_0(\cdot,\epsilon), b_1(\cdot,\epsilon), \sigma_0(\cdot,\epsilon) \in G^3$.

Now, since $b_0(x,\epsilon) \to b_0(x,0)$ (and similarly for b_1, σ_0), and the limits are continuous, by Lem. 1, 2, 12, 13,, all conditions of Thm. 1 are satisfied, and hence we conclude that for each $g \in G^3$, we have,

$$\max_{k=0,\ldots,N} |\mathbb{E}g(X^\epsilon_{k\eta}) - \mathbb{E}g(x_k)| \leq C(\eta^2 + \rho(\epsilon))$$

where C is independent of η and ϵ and $\rho(\epsilon) \to 0$ as $\epsilon \to 0$. Moreover, since $b_0(x,\epsilon) \to b_0(x,0)$ (and similarly for b_1, σ_0) uniformly on compact sets, we may apply Thm. 7 to conclude that

$$\sup_{t\in[0,T]} \mathbb{E}|X^\epsilon_t - X_t|^2 \to 0 \text{ as } \epsilon \to 0$$

Thus, we have

$$\begin{aligned}
&|\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| \\
\leq &|\mathbb{E}g(X^\epsilon_{k\eta}) - \mathbb{E}g(x_k)| + |\mathbb{E}g(X^\epsilon_{k\eta}) - \mathbb{E}g(X_{k\eta})| \\
\leq &C(\eta^2 + \rho(\epsilon)) + \left(\mathbb{E}|X^\epsilon_{k\eta} - X_{k\eta}|^2\right)^{1/2} \\
&\times \left(\int_0^1 \mathbb{E}|\nabla^2 g(\lambda X^\epsilon_{k\eta} + (1-\lambda)X_{k\eta})|^2 d\lambda\right)^{1/2}
\end{aligned}$$

Using Thm. 6 and assumption that $\nabla^2 g \in G$, the last expectation is finite and hence taking the limit $\epsilon \to 0$ yields our result. □

By going for a lower order approximation, we of course have the following:

Corollary 1. *Assume the same conditions as in Thm. 2, except that we replace (i) with*

(i)' $f \equiv \mathbb{E}f_\gamma$ is continuously differentiable, and $f \in G^3_w$.

Define $\{X_t : t \in [0,T]\}$ as the stochastic process satisfying the SDE

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta}\Sigma(X_t)^{1/2}dW_t \qquad X_0 = x_0 \tag{18}$$

with $\Sigma(x) = \mathbb{E}(\nabla f_\gamma(x) - \nabla f(x))(\nabla f_\gamma(x) - \nabla f(x))^T$. Then, $\{X_t : t \in [0,T]\}$ is an order-1 weak approximation of the SGD, i.e. for each $g \in G^2$, there exists a constant $C > 0$ independent of η such that

$$\max_{k=0,\ldots,N} |\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| \leq C\eta$$

Remark. In the above results, the most restrictive condition is probably the Lipschitz condition on ∇f_γ. Such Lipschitz conditions are important to ensure that the SMEs admit unique strong solutions and the SGA having uniformly bounded moments. Note that following similar techniques in SDE analysis (e.g. [Kloeden and Platen(2011)]), these global conditions may be relaxed to their respective local versions if we assume in addition a uniform global linear growth condition on ∇f_γ. Finally, for applications, typical loss functions have inward pointing gradients for all sufficiently large x, meaning that the SGD iterates will be uniformly bounded almost surely. Thus, we may simply modify the loss functions for large x (without affecting the SGA iterates) to satisfy the conditions above.

Remark. The constant C does not depend on η, but as evidenced in the proof of the theorem, it generally depends on g, T, d and the various Lipschitz constants. For the fairly general situation we are consider, we do not derive tight estimates of these dependencies.

4.3 SME for stochastic gradient descent with momentum

Let us discuss the corresponding SME for a popular variant of the SGD called the *momentum SGD* (MSGD). The momentum SGD augments the usual SGD iterations with a "memory" term. In the usual form, we have the iterations

$$\begin{aligned}\widehat{v}_{k+1} &= \widehat{\mu}\widehat{v}_k - \widehat{\eta}\nabla f_{\gamma_k}(x_k)\\ x_{k+1} &= x_k + \widehat{v}_{k+1}\end{aligned}$$

where $\widehat{\mu} \in (0,1)$ (typically close to 1) is called the *momentum parameter* and $\widehat{\eta}$ is the learning rate. Let us consider a rescaled version of the above that is easier to analyze via continuous-time approximations. We redefine

$$\eta := \sqrt{\widehat{\eta}}, \qquad v_k := \widehat{v}_k/\sqrt{\widehat{\eta}}, \qquad \mu := (1-\widehat{\mu})/\sqrt{\widehat{\eta}} \tag{19}$$

to obtain

$$\begin{aligned}v_{k+1} &= v_k - \mu\eta v_k - \eta\nabla f_{\gamma_k}(x_k)\\ x_{k+1} &= x_k + \eta v_{k+1}.\end{aligned} \tag{20}$$

In view of the rescaling, the range of momentum parameters we consider becomes $\mu \in (0, \eta^{-1/2})$, which we may replace by $(0,\infty)$ for simplicity.

Let us now derive the SME satisfied by the iterations (20). Observe that this is again a special case of (14) with x now replaced by (v,x) and

$$h(v,x,\gamma,\eta) = (-\mu v - \nabla f_\gamma(x), v - \eta\mu v - \eta\nabla f_\gamma(x))$$

In view of Thm. 3 and the results in Sec. 4.2, in order to derive the SMEs we simply match moments up to order 3. As in Sec. 4.2, let us define the one step difference

$$\Delta(v,x) := (v_1^{v,x,0} - v, x_1^{v,x,0} - x) \tag{21}$$

The following moment expansions are immediate.

Lemma 3. *Let $\Delta(x,v)$ be defined as in (21). We have*

(i) $\mathbb{E}\Delta_{(i)}(v,x) = \eta(-\mu v_{(i)} - \partial_{(i)}f(x), v) + \eta^2(0, -\mu v_{(i)} - \partial_{(i)}f(x))$,

(ii) $\mathbb{E}\Delta_{(i)}(v,x)\Delta_{(j)}(v,x) =$
$$\eta^2\begin{pmatrix}\mu^2 v_{(i)}v_{(j)} + \mu v_{(i)}\partial_{(j)}f(x) + \mu v_{(j)}\partial_{(i)}f(x) & \\ \quad +\Sigma(x)_{(i,j)} + \partial_{(i)}\partial_{(j)}f(x) & -\mu v_{(i)}v_{(j)} - v_{(i)}\partial_{(j)}f(x)\\ -\mu v_{(i)}v_{(j)} - v_{(j)}\partial_{(i)}f(x) & v_{(i)}v_{(j)}\end{pmatrix}$$
$+ \mathcal{O}(\eta^3)$,

(iii) $\mathbb{E}\prod_{j=1}^3 |\Delta_{(i_j)}(v,x)| = \mathcal{O}(\eta^3)$,

where $\Sigma(x) := \mathbb{E}(\nabla f_\gamma(x) - \nabla f(x))(\nabla f_\gamma(x) - \nabla f(x))^T$.

Proof. The proof follows from direct calculation of the moments. □

Hence, proceeding exactly as in Sec. 4.2 and using Lem.1, 3, we see that we may set

$$
\begin{aligned}
b_0(v,x) &= (-\mu v - \nabla f(x), v) \\
b_1(v,x) &= -\frac{1}{2}\left(\mu[\mu v + \nabla f(x)] - \nabla^2 f(x) v, \mu v + \nabla f(x)\right) \\
\sigma_0(v,x) &= \begin{pmatrix} \Sigma(x)^{1/2} & 0 \\ 0 & 0 \end{pmatrix}
\end{aligned}
$$

in order to match the moments. By similar mollification and limiting arguments as in Thm. 2, we arrive at the following approximation theorem, where we can see that the SME for MSGD takes the form of a Langevin equation.

Theorem 3. *Assume the same conditions as in Thm. 2. Let $\mu > 0$ be fixed and define $\{V_t, X_t : t \in [0,T]\}$ as the stochastic process satisfying the SDE*

$$
\begin{aligned}
dV_t &= -[(\mu I + \frac{1}{2}\eta[\mu^2 I - \nabla^2 f(X_t)])V_t + (1 + \frac{1}{2}\eta\mu)\nabla f(X_t)]dt + \sqrt{\eta}\Sigma(X_t)^{1/2} dW_t \quad V_0 = v_0, \\
dX_t &= [(1 - \frac{1}{2}\eta\mu)V_t - \frac{1}{2}\eta\nabla f(X_t)]dt \quad X_0 = x_0
\end{aligned} \tag{22}
$$

with $\Sigma(x)$ as defined in Thm. 2. Then, $\{(V_t, X_t) : t \in [0,T]\}$ is an order-2 weak approximation of the MSGD.

Moreover, if we relax the assumptions to Cor. 1, we have the order-1 weak approximation

$$
\begin{aligned}
dV_t &= -[\mu V_t + \nabla f(X_t)]dt + \sqrt{\eta}\Sigma(X_t)^{1/2} dW_t \quad V_0 = v_0, \\
dX_t &= V_t dt \quad X_0 = x_0
\end{aligned} \tag{23}
$$

Note that by inverting the scaling (19), the order-1 SME (23) is the formal equation derived in [Li et al.(2015)].

4.4 SME for a momentum variant: Nesterov accelerated gradient

It follows from the calculation above that we can also obtain the SME for the stochastic gradient version of the Nesterov accelerated gradient (NAG) method [Nesterov(1983)], which we refer to as SNAG. In the non-stochastic case, the NAG method has been analyzed using the ODE approach [Su et al.(2014)]. Therefore, the derivations in this section can be viewed as a stochastic parallel. The NAG method is sometimes used with stochastic gradients, and hence it is useful to analyze its properties in this setting and compare it to MSGD.

The unscaled NAG iterations are

$$
\begin{aligned}
\widehat{v}_{k+1} &= \widehat{\mu}_k \widehat{v}_k - \widehat{\eta}\nabla f_{\gamma_k}(x_k + \widehat{\mu}_k \widehat{v}_k) \\
x_{k+1} &= x_k + \widehat{v}_{k+1}
\end{aligned}
$$

with $\widehat{v}_0 = 0$, which differs from the momentum iterations as the gradient is now evaluated at a "predicted" position $x_k + \widehat{\mu}_k \widehat{v}_k$, instead of the original position x_k. Moreover, the momentum parameter $\widehat{\mu}_k$ is now allowed to vary as k increases, and in fact, the usual choice of

$$
\widehat{\mu}_k = \frac{k-1}{k+2} \tag{24}
$$

this has important links to stability and acceleration in the deterministic case [Nesterov(1983), Su et al.(2014)]. In particular, it achieves $\mathcal{O}(1/k^2)$ convergence rate for general convex functions. On the other hand, a constant $\widehat{\mu}_k$ is suggested for strongly convex functions [Nesterov(2013)]. In the following, we shall first consider the case of constant momentum parameter with $\widehat{\mu}_k \equiv \widehat{\mu}$, and then the choice (24) subsequently.

Constant momentum. Using the same rescaling in (19), we have

$$\begin{aligned} v_{k+1} &= v_k - \mu\eta v_k - \eta\nabla f_{\gamma_k}(x_k + \eta(1-\mu\eta)v_k) \\ x_{k+1} &= x_k + \eta v_{k+1}. \end{aligned} \tag{25}$$

which is again (14) with

$$h(v, x, \gamma, \eta) = (-\mu v - \nabla f_\gamma(x + \eta(1-\mu\eta)v), v - \eta\mu v - \eta\nabla f_\gamma(x + \eta(1-\mu\eta)v))$$

Hence, we have the following moment expansion.

Lemma 4. *Let* $\Delta(x, v) := (v_1^{v,x,0} - v, x_1^{v,x,0} - x)$. *We have*

(i) $\mathbb{E}\Delta_{(i)}(v,x) = \eta(-\mu v_{(i)} - \partial_{(i)}f(x), v)$
$+ \eta^2(\partial_{(i)}\partial_{(j)}f(x)v_{(j)}, -\mu v_{(i)} - \partial_{(i)}f(x+v)) + \mathcal{O}(\eta^3)$,

(ii) $\mathbb{E}\Delta_{(i)}(v,x)\Delta_{(j)}(v,x) =$
$\eta^2 \begin{pmatrix} \mu^2 v_{(i)}v_{(j)} + \mu v_{(i)}\partial_{(j)}f(x+v) + \mu v_{(j)}\partial_{(i)}f(x+v) & \\ \quad +\Sigma(x+v)_{(i,j)} + \partial_{(i)}\partial_{(j)}f(x+v) & -\mu v_{(i)}v_{(j)} - v_{(i)}\partial_{(j)}f(x+v) \\ -\mu v_{(i)}v_{(j)} - v_{(j)}\partial_{(i)}f(x+v) & v_{(i)}v_{(j)} \end{pmatrix}$
$+ \mathcal{O}(\eta^3)$,

(iii) $\mathbb{E}\prod_{j=1}^3 |\Delta_{(i_j)}(v,x)| = \mathcal{O}(\eta^3)$,

where $\Sigma(x) := \mathbb{E}(\nabla f_\gamma(x) - \nabla f(x))(\nabla f_\gamma(x) - \nabla f(x))^T$.

Proof. The proof follows from direct calculation of the moments and Taylor's expansion. □

Hence, we may match moments by setting

$$\begin{aligned} b_0(v,x) &= (-\mu v - \nabla f(x), v) \\ b_1(v,x) &= -\frac{1}{2}\left(\mu[\mu v + \nabla f(x)] + \nabla^2 f(x)v, \mu v + \nabla f(x)\right) \\ \sigma_0(v,x) &= \begin{pmatrix} \Sigma(x)^{\frac{1}{2}} & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

from which we obtain the following approximation theorem for SNAG.

Theorem 4. *Assume the same conditions as in Thm. 3. Define* $\{V_t, X_t : t \in [0,T]\}$ *as the stochastic process satisfying the SDE*

$$\begin{aligned} dV_t &= -[(\mu I + \frac{1}{2}\eta[\mu^2 I + \nabla^2 f(X_t)])V_t + (1 + \frac{1}{2}\eta\mu)\nabla f(X_t)]dt + \sqrt{\eta}\Sigma(X_t)^{1/2}dW_t \quad V_0 = v_0, \\ dX_t &= [(1 - \frac{1}{2}\eta\mu)V_t - \frac{1}{2}\eta\nabla f(X_t)]dt \quad X_0 = x_0 \end{aligned} \tag{26}$$

with Σ *as defined in Thm. 3. Then,* $\{(V_t, X_t) : t \in [0,T]\}$ *is an order-2 weak approximation of SNAG. Moreover, the same order-1 weak approximation of MSGD in* (23) *holds for the SNAG.*

The result above shows that for constant momentum parameters, the modified equations for MSGD and the SNAG are equivalent at leading order, but differ when we consider the second order modified equation. Let us now discuss the case where the momentum parameter is allowed to vary.

Varying momentum. Now let us take $\widehat{\mu}$ as in (24). Then, using the same rescaling arguments, we arrive at

$$\begin{aligned} v_{k+1} &= v_k - \mu_k \eta v_k - \eta \nabla f_{\gamma_k}(x_k + \eta(1-\mu_k\eta)v_k) \\ x_{k+1} &= x_k + \eta v_{k+1}. \end{aligned} \tag{27}$$

with $\mu_k = 3/(2\eta + k\eta)$. Now, in order to apply our theoretical results to deduce the SME, simply notice that we may introduce an auxiliary scalar variable

$$z_{k+1} = z_k + \eta, \qquad z_0 = 0$$

Then, $\mu_k = 3/(2\eta + z_k)$, and hence all terms are now not explicitly k-independent, thus we may proceed formally as in the previous sections to arrive at the order-1 SME for SNAG with varying momentum

$$\begin{aligned} dV_t &= -[\frac{3}{t}V_t + \nabla f(X_t)]dt + \sqrt{\eta}\Sigma(X_t)^{1/2}dW_t \quad V_0 = 0, \\ dX_t &= V_t dt \quad X_0 = x_0 \end{aligned} \tag{28}$$

This result is formal because the term $3/t$ does not satisfy our global Lipschitz conditions, unless we restrict our interval to some $[t_0, T]$ with $t_0 > 0$, in which case the above result becomes rigorous. Alternatively, some limiting arguments have to be used to establish well-posedness of the equation on $[0, T]$ individually. We shall omit these analyses in the current paper, and only consider (28) on some interval $[t_0, T]$, where initial conditions are then replaced by (v_{t_0}, x_{t_0}). As a point of comparison, (28) reduces to the ODE derived in [Su et al.(2014)] if $\Sigma(x) \equiv 0$ (i.e. the gradients are non-stochastic).

5 Applications of the SMEs to the analysis of SGA

In this section, we apply the SME framework developed to analyze the dynamics of the three stochastic gradient algorithm variants discussed above, namely SGD, MSGD and SNAG. We shall focus on simple but non-trivial models where to a large extent, analytical computations using SME are tractable, giving us key insights into the algorithms that are otherwise difficult to obtain without appealing to the continuous formalism presented in this paper. We consider primarily the following model:

Model: Let $H \in \mathbb{R}^{d\times d}$ be a symmetric, positive definite matrix. Define the sample objective

$$\begin{aligned} f_\gamma(x) &:= \frac{1}{2}(x-\gamma)^T H(x-\gamma) - \frac{1}{2}\mathrm{Tr}(H) \\ \gamma &\sim \mathcal{N}(0, I) \end{aligned} \tag{29}$$

which gives the total objective $f(x) \equiv \mathbb{E}f_\gamma(x) = \frac{1}{2}x^T H x$.

5.1 SME analysis of SGD

We first derive the SME associated with (29). For simplicity, we will only consider the order-1 SME (18). A direct computation shows that $\Sigma(x) = H^2$ and so the SME for SGD applied to model (29) is

$$dX_t = -HX_t dt + \sqrt{\eta} H dW_t$$

This is a multi-dimensional Ornstein-Uhlenbeck (OU) process and admits the explicit solution

$$X_t = e^{-tH}\left(x_0 + \sqrt{\eta}\int_0^t e^{sH} H dW_s\right)$$

Observe that for each $t \geq 0$, the distribution of X_t is Gaussian. Using Itô's isometry, we then deduce the dynamics of the objective function

$$\begin{aligned}\mathbb{E}f(X_t) =& \frac{1}{2}x_0^T H e^{-2tH} x_0 + \frac{1}{2}\eta\int_0^t \mathrm{Tr}(H^3 e^{-2(t-s)H}) ds \\ =& \frac{1}{2}x_0^T H e^{-2tH} x_0 + \frac{1}{4}\eta\sum_{i=1}^n \lambda_i^2(H)(1 - e^{-2t\lambda_i(H)})\end{aligned} \tag{30}$$

The first term decays linearly with asymptotic rate $2\lambda_d(H)$, and the second term is induced by noise, and its asymptotic value is proportional to the learning rate η. This is the well-known two-phase behavior of SGD under constant learning rates: an initial descent phase induced by the deterministic gradient flow and an eventual fluctuation phase dominated by the variance of the stochastic gradients. In this sense, the SME makes the same predictions, and in fact we can see that it approximates the SGD iterations well as η decreases (Fig. 1(a)), according to the rates we derived in Thm. 2 and Cor. 1.

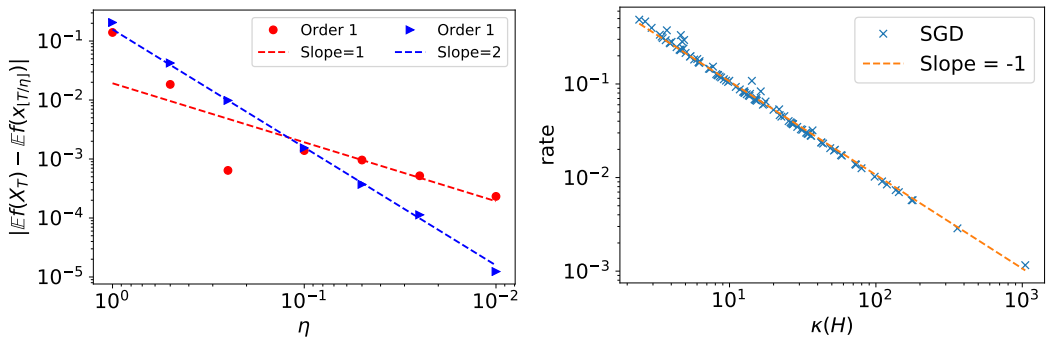
Moreover, notice that by the identification $t = k\eta$ (k is the SGD iteration number), the SME analysis tells us that the asymptotic linear convergence rate (in k, i.e. rate $\sim -\log[\mathbb{E}f(x_k)]/k$) in the descent phase of the SGD is $2\lambda_d(H)\eta$. For numerical stability (even in the non-stochastic case), we usually require $\eta \propto 1/\lambda_1(H)$, thus the maximal descent rate is inversely proportional to the condition number $\kappa(H) = \lambda_1(H)/\lambda_d(H)$. We validate this observation by generating a collection of H's with varying condition numbers and applying SGD with $\eta \propto 1/\lambda_1(H)$. In Fig 1(b), we plot the initial descent rates versus the condition number of H and we observe that we indeed have rate $\propto \kappa(H)^{-1}$.

Alternate model. Now, we consider a slight variation of the model (29). The goal is show that the dynamics of SGD (and the corresponding SME) is not always Gaussian-like and thus using the OU process to model the SGD is not always valid. Given the same positive-definite matrix H, we diagonalize it in the form $H = QDQ^T$ where Q is an orthogonal matrix and D is a diagonal matrix of eigenvalues. We then define the sample objective

$$\begin{aligned}&f_\gamma(x) := \frac{1}{2}(Q^T x)^T[D + \mathrm{diag}(\gamma)](Q^T x) \\ &\gamma \sim \mathcal{N}(0, I)\end{aligned} \tag{31}$$

which gives the same total objective $f(x) \equiv \mathbb{E}f_\gamma(x) = \frac{1}{2}x^T H x$. However, we have a different expression for $\Sigma(x)$

$$\Sigma(x) = Q\mathrm{diag}(Qx)^2 Q^T$$



(a) (b)

Figure 1. SME prediction vs SGD dynamics. (a) SME as a weak approximation of the SGD. We compute the weak error with test function g equal to f (see Thm. 2). As predicted by our analysis, the order-2 SME (17) (order-1 SME (18)) should give a slope $= 2$ (1) decrease in error as η decreases (note that the x-axis is flipped). The SME solution is computed using an exact formula derived by the application of Itô isometry and the SGD expectation is averaged over 1e6 runs. We took $T = 2.0$. We see that the predictions of Thm. 2 and Cor.1 hold. (b) Descent rate vs condition number. H is generated with different condition numbers, and the resulting descent rate of SGD is approximately $\propto \kappa(H)^{-1}$, as predicted by the SME.

which gives the SME

$$dX_t = -HX_t dt + \sqrt{\eta} Q |\operatorname{diag}(Q^T x)| Q^T dW_t$$
$$\overset{\text{in distribution}}{=} -HX_t dt + \sqrt{\eta} Q \operatorname{diag}(Q^T x) Q^T dW_t$$

We can rewrite the above as

$$dX_t = -HX_t dt + \sqrt{\eta} \sum_{l=1}^{d} Q^{(l)} X_t dW_{(l),t}$$

where $Q^{(l)} = Q \operatorname{diag}(Q_{(l,\cdot)}) Q^T$ and $Q_{(l,\cdot)}$ denotes the l^{th} row of Q. By observing that every pair of $\{H, Q^{(1)}, \ldots, Q^{(d)}\}$ commute, we have the explicit solution

$$X_t = e^{-\frac{1}{2}\eta t + \sqrt{\eta}\sum_{l=1}^{d} Q^{(l)} W_{(l),t}} e^{-Ht} x_0$$

which is a multi-dimensional Black-Scholes [Black and Scholes(1973)] type of stochastic process. In particular, the distribution is not Gaussian of any $t > 0$. Nevertheless, we may take expectation to obtain

$$\mathbb{E}f(X_t) = \frac{1}{2} e^{\eta t} x_0^T H e^{-2Ht} x_0$$

This immediately implies the following interesting behavior: if $\eta < 2\lambda_d(H)$, then $2H - \eta I$ is positive definite and so $\mathbb{E}f(X_t) \to 0$ exponentially at constant, non-zero η; Otherwise, depending on initial condition x_0, the objective may not converge to 0. In particular, if $\eta > 2\lambda_d(H)$ (which happens quite often if the condition number of H is large) and x_0 is in general position, then we have asymptotic exponential divergence. This is a variance-induced divergence typically observed in

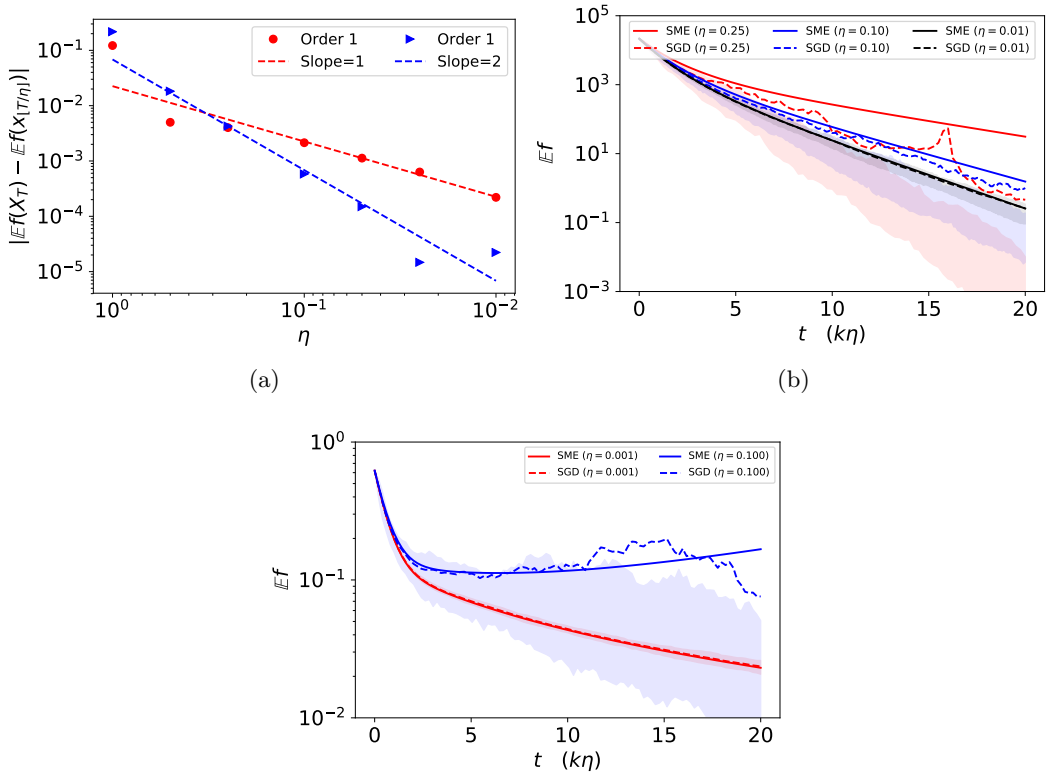


Figure 2. SME prediction vs SGD dynamics for the model variant (31). (a) Order of convergence of the SME to the SGD. We use the same setup as in Fig. 1(a). Observe that our analysis again predicts the correct rate of weak error decay as η decreases. (b) SGD paths vs order-1 SME prediction. Solid lines are SME exact solutions and dotted lines are means of SGD paths over 500 runs, and the 25-75 percentiles are shaded. We observe convergence of $\mathbb{E}f$ at constant η, and that the sample mean is dominated by few large values, as observed by the deviation of the percentiles from the mean. (b) Variance-induced explosion. As predicted by the SME analysis, if $\eta > 2\lambda_d(H)$ (Here, $\lambda_d(H) = 0.01$), variance-induced instability sets in.

Black-Scholes and geometric Brownian motion type of stochastic processes. The term "variance-induced" is important here since the deterministic part of the evolution equation is mean-reverting and in fact is identical to the stable OU process studied earlier. In Fig. 2(a), (b), we show the correspondence of the SME findings with the actual dynamics of the SGD iterations. In particular, we see in Fig. 2(c) that for small η, we have exponential convergence of the SGD at constant learning rates, whereas for $\eta > 2\lambda_d(H)$, the SGD iterates start to oscillate wildly and its mean value is dominated by few large values and diverges approximately at the rate predicted by the SME. Note that this divergence is predicted to be at a finite η, and from the theory developed so far we cannot conclude that the SME approximation always holds accurately at this regime (but the approximation is guaranteed for η sufficiently small). Nevertheless, we observe at least in this model that the variance-induced divergence of the SGD happens as predicted by the SME.

5.2 SME analysis of MSGD

Let us now use the SME to analyze MSGD applied to model (29). We have shown earlier that $\Sigma(x) = H$. Thus, according to Thm. 3, the order-1 SME for MSGD is

$$\begin{aligned} dV_t &= -[\mu V_t + HX_t]dt + \sqrt{\eta} H dW_t, \\ dX_t &= V_t dt, \end{aligned} \tag{32}$$

with $X_0 = x_0$ and $V_0 = 0$. If we set $Y_t := (V_t, X_t) \in \mathbb{R}^{2d}$, U_t a $2d$-dimensional Brownian motion with first d coordinates equal to W_t, and define block matrices

$$A := \begin{pmatrix} \mu I & H \\ -I & 0 \end{pmatrix}, \qquad B := \begin{pmatrix} H & 0 \\ 0 & 0 \end{pmatrix} \tag{33}$$

we can then write (32) as

$$dY_t = -AY_t + \sqrt{\eta} B dU_t, \qquad Y_0 = (0, x_0)$$

which admits the explicit solution

$$Y_t = e^{-At}\left(Y_0 + \sqrt{\eta}\int_0^t e^{As} B dU_s.\right)$$

By Itô's isometry, we have

$$\mathbb{E}f(X_t) = \frac{1}{2}\left[|\operatorname{diag}(0,H)^{1/2} e^{-At} Y_0|^2 + \eta \int_0^t |\operatorname{diag}(0,H)^{1/2} e^{-(t-s)A} B|^2 ds\right] \tag{34}$$

One can see immediately that a similar two-phase behavior is present, but the property of the descent phase now hinges on the spectral properties of the matrix A (instead of H). Before proceeding, we first observe that the eigenvalues of A can be written as

$$\lambda(A) := \{\Lambda_+, \Lambda_-\}, \qquad \Lambda_{\pm,i} = \frac{1}{2}\Big(\mu \pm \sqrt{\mu^2 - 4\lambda_i(H)}\Big), \qquad i = 1, 2, \ldots, d \tag{35}$$

In particular, $\Re\lambda_i(A) > 0$ for all i as long as $\mu > 0$. We also need the following simple result concerning the decay of the norm of e^{-tA} if all eigenvalues of A have positive real part.

Lemma 5. *Let A be a real square matrix such that all eigenvalues have positive real part. Then,*

(i) For each $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ independent of t but depends on ϵ, such that

$$|e^{-tA}| \le C_\epsilon e^{-t(\min_i \Re\lambda_i(A) - \epsilon)}$$

(ii) If in addition A is diagonalizable, then there exists a constant $C > 0$ independent of t such that

$$|e^{-tA}| \le C e^{-t \min_i \Re\lambda_i(A)}$$

Proof. See Appendix E. □

With the above results, we can now characterize the decay of the objective under momentum SGD. From expression (35), we see that as long as $\mu^2 \neq 4\lambda_i$ for any $i = 1, \ldots, d$, A has $2d$ distinct eigenvalues and is hence diagonalizable. We shall hereafter assume that μ is in general position such that this is the case. Using Lem. 5 and expression (34), we arrive at the estimate

$$\mathbb{E}f(X_t) \leq \frac{1}{2}C^2|x_0|^2\lambda_1(H)e^{-2t\min_i \Re\lambda_i(A)} + \frac{1}{2}\frac{\eta C^2\lambda_1(H)^3}{\min_i \Re\lambda_i(A)}(1 - e^{-2t\min_i \Re\lambda_i(A)}) \tag{36}$$

This result tells us that the convergence rate of the descent phase is now controlled by the minimum real part of the eigenvalues of A, instead of the minimum eigenvalue of H. In particular, we achieve the best linear convergence rate by maximizing the smallest real part of the eigenvalues of A. This leads to the following optimization problem for the optimal convergence rate:

$$\sup_{\mu\in(0,\infty)} \min_{i=1,\ldots,d} \min_{s\in\{+1,-1\}} \left\{\Re\left[\mu + s\sqrt{\mu^2 - 4\lambda_i(H)}\right]\right\}$$

Since H is positive definite, the supremum is attained at $\mu^* = 2\sqrt{\lambda_d(H)}$ with the rate also equal to $2\sqrt{\lambda_d(H)}$. However, note that if we take $\mu = \mu^*$ exactly, one can check that A is no longer diagonalizable and by Lem. 5, the rate is slightly diminished, thus technically we can take μ as close to μ^* as we like (i.e. the rate is as close to $2\sqrt{\lambda_d(H)}$ as we like), but exact equality is not technically deducible from current results. In Fig. 3(c), we demonstrate the optimal choice of μ and its effect on the convergence rate. Moreover, observe that as μ increases, the number of complex eigenvalues start to decrease, and the magnitudes of the imaginary parts of the complex eigenvalues also decrease. This signifies that increasing μ causes oscillations to decreases in magnitude and frequency. This is again corroborated by numerical experiments (Fig. 3(c)).

Another interesting observation is that by the identification $t = \eta k$, the descent rate (in terms of k) is $2\sqrt{\lambda_d(H)}\eta$. As before, if we choose the maximal stable learning rate we would have $\widehat{\eta} \propto 1/\lambda_1(H)$ ($\widehat{\eta} = \eta^2$ according to the scaling introduced in (19)). Thus, for the MSGD iterates we have its descent rate $\propto \kappa(H)^{-1/2}$, which is a huge improvement over SGD, whose rate is $\propto \kappa(H)^{-1}$, especially for badly conditioned matrices where $\kappa(H) \gg 1$. In Fig. 3(d), we plot the MSGD initial descent rates for varying condition numbers of H. Again, we observe that the SME analysis gives the correct characterization of the precise dynamics and recovers the square-root relationship with condition number.

Finally, let us discuss the effect of adding momentum to the asymptotic fluctuations due to noisy gradients. Note that it is not correct to conclude, using Eq. (36), that taking $\mu \approx \mu^*$ also gives the lowest fluctuations. This is because the constant C depends on μ as well, as is evidenced in the proof of Lem. 5, which shows that C depends on the conditioning of the eigenvector matrix of A. To proceed, we do not use the bounds (36). Instead, we explicitly diagonalize A and after some computations, we arrive at the exact expression for $\mathbb{E}f(X_t)$

$$\mathbb{E}f(X_t) = \frac{1}{2}|\operatorname{diag}(0, H)^{1/2}e^{-At}Y_0|^2 \tag{37}$$

$$+ \frac{1}{2}\eta\sum_{i=1}^{d}\frac{\lambda_i^3}{|\mu^2 - 4\lambda_i|}\left[\frac{1 - e^{-2t\Re\Lambda_{+,i}}}{2\Re\Lambda_{+,i}} + \frac{1 - e^{-2t\Re\Lambda_{-,i}}}{2\Re\Lambda_{-,i}} - 2R(t, \mu, \lambda_i(H))\right] \tag{38}$$

where

$$R(t, \mu, \lambda) = \begin{cases} \frac{1-e^{-t\mu}}{\mu} & \mu \geq 2\sqrt{\lambda} \\ \frac{\mu + \sqrt{4\lambda - \mu^2}e^{-\mu t}\sin(t\sqrt{4\lambda-\mu^2}) - \mu e^{-\mu t}\cos(t\sqrt{4\lambda-\mu^2})}{4\lambda} & \mu < 2\sqrt{\lambda} \end{cases} \tag{39}$$

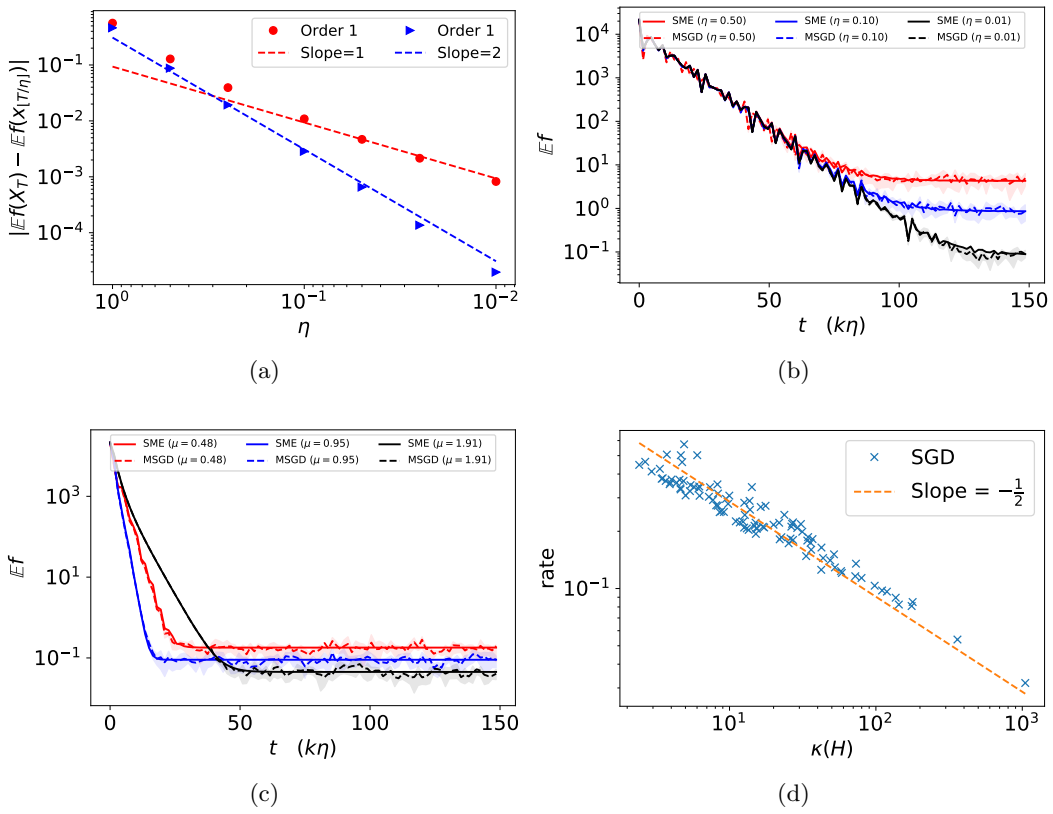


Figure 3. SME prediction vs MSGD dynamics. (a) and (b) SME vs MSGD dynamics at $\mu = 0.1$ for different learning rates η. As before, the SME prediction gets better as η decreases according to the predicted order. Notice also the presence of oscillations, due to the complex eigenvalues of A. (c) Optimal descent rate of the SGD is achieved by the SME prediction $\mu = \mu^*$, which is 0.95 in this case. Notice that exactly as predicted by the SME, increasing μ decreases the oscillation frequency and magnitude (due to having fewer complex eigenvalues and smaller imaginary parts), as well as the asymptotic fluctuations (due to formula (37)). (d) Descent rate vs condition number. H is generated with different condition numbers, and the descent rate of MSGD is $\propto \kappa(H)^{-1/2}$, as predicted by the SME, which for badly conditioned H gives a large improvement.

In particular, the asymptotic loss value induced by noise is

$$\lim_{t\to\infty} \mathbb{E}f(X_t) = \frac{1}{2}\eta \sum_{i=1}^{d} \frac{\lambda_i(H)^3}{|\mu^2 - 4\lambda_i(H)|} \left[\frac{1}{2\Re\Lambda_{+,i}} + \frac{1}{2\Re\Lambda_{-,i}} - 2\min\left\{ \frac{\mu}{4\lambda_i(H)}, \frac{1}{\mu} \right\} \right] \tag{40}$$

Observe that this function (in fact, each term in the sum) is monotone-decreasing in μ, and for $\mu \ll 1$ it scales like μ^{-1}, and for $\mu \gg 1$ it scales like μ^{-3}. Thus, increasing the momentum parameter decreases the asymptotic noise in the iterates, i.e. decreases the asymptotic value of $\mathbb{E}f$, which should be 0 in the absence of noise. This again agrees with the actual MSGD dynamics (Fig. 3(b)). Consequently, to obtain "optimal" tradeoff between descent and noise, we would like a momentum schedule that equals μ^* in the descent phase and increases to infinity (in the original scaling this corresponds to $\widehat{\mu} \to 0$) as we approach the optimum. Finding this optimal schedule can be cast as an optimal control problem [Li et al.(2017)], and a rigorous investigation of these approaches will be considered in subsequent work.

5.3 SME analysis of SNAG

Finally, let us see what we can say, using the SME approach, about the difference between MSGD and SNAG in this stochastic setting. Let us first consider the case of constant momentum. From Thm. 4, we know that the order-1 SMEs are identical, so we must consider higher order SMEs. A straightforward computation yields the following order-2 SMEs for MSGD and SNAG (again we let $Y_t = (V_t, X_t)$)

$$\begin{aligned} &\text{MSGD:} && dY_t = -A_1 Y_t + \sqrt{\eta} B dU_t, && Y_0 = (0, x_0), \\ &\text{SNAG:} && dY_t = -A_2 Y_t + \sqrt{\eta} B dU_t, && Y_0 = (0, x_0) \end{aligned}$$

where $A_i = A + \frac{1}{2}\eta E_i$ with A, B as defined in (33) and

$$E_1 := \begin{pmatrix} \mu^2 I - H & \mu H \\ \mu I & H \end{pmatrix}, \qquad E_2 := \begin{pmatrix} \mu^2 I + H & \mu H \\ \mu I & H \end{pmatrix}$$

From the analysis in Sec. 4.3, the descent rate is dominated by the minimal real parts of the eigenvalues of A_i, which are respectively

$$\begin{aligned} \lambda(A_1) &= \left\{ \frac{1}{4}\left(\mu(\eta\mu + 2) \pm \sqrt{\mu^2(\eta\mu+2)^2 + 4\eta^2\lambda_i(H)^2 - 8\lambda_i(H)(\eta\mu+2)} \right), \quad i = 1, \dots, d \right\} \\ \lambda(A_2) &= \left\{ \frac{1}{4}\left(\mu(\eta\mu + 2) + 2\eta\lambda_i(H) \pm \sqrt{\eta\mu+2}\sqrt{\mu^2(\eta\mu+2) + 4\lambda_i(H)(\eta\mu-2)} \right), \quad i = 1, \dots, d \right\} \end{aligned}$$

We observe that for small μ (i.e. $\widehat{\mu} \approx 1$ in the usual MSGD scaling), the terms in square-roots are negative and hence for the same small μ, the convergence rate of SNAG is $\frac{1}{2}\eta\lambda_d(H)$ larger than that of MSGD. This says in particular that for H with larger $\lambda_d(H)$, the acceleration is more pronounced. Moreover, recall that the asymptotic fluctuations is given by

$$\eta \lim_{t\to\infty} \int_0^t |\operatorname{diag}(0, H)^{1/2} e^{-(t-s)(A + \frac{1}{2}\eta E_i)} B|^2 ds$$

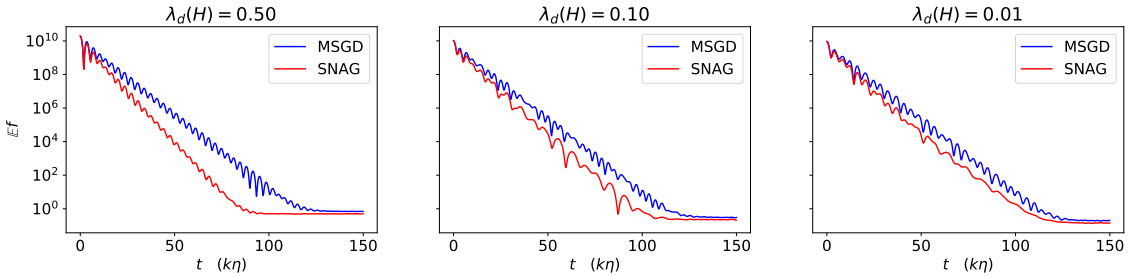
Without performing tedious computations, we can see that since $E_2 - E_1$ is positive definite, the exponential for the SNAG case decays more rapidly, and hence the eventual fluctuations are expected to be lower. These observations from the SME are again consistent with the behavior of their SGA counter-parts, as shown in Fig. 4(a). On the other hand, if we pick μ for each case by separately maximizing the smallest real part of the eigenvalues (as in Sec. 4.3), we obtain similar convergence rates up to η^2. In other words, if we tune μ well, there would be no difference between MSGD and SNAG in terms of descent rate (Fig. 4(b)).

Now, let us discuss the varying momentum case. According to (28), for some small $t_0 > 0$ we have the order-1 SME for $t \in [t_0, T]$

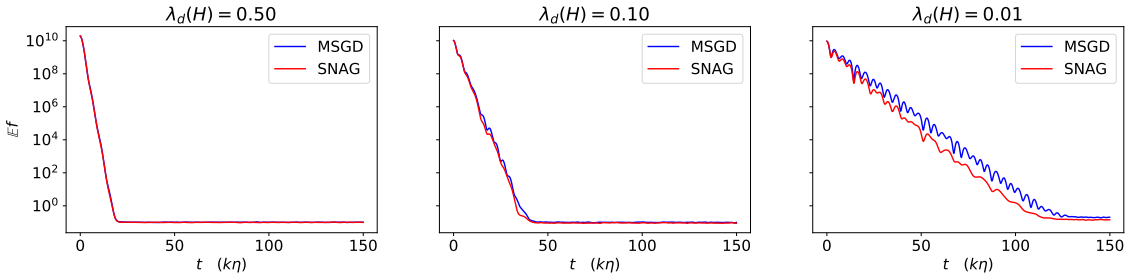
$$dY_t = -A_t Y_t + \sqrt{\eta} B dU_t, \qquad Y_{t_0} = (v_{t_0}, x_{t_0}) \qquad A_t := \begin{pmatrix} \frac{3}{t} I & H \\ -I & 0 \end{pmatrix}$$

and B is defined as in (33). This admits the explicit solution

$$Y_t = e^{-(t-t_0)\widetilde{A}_t} \left(Y_{t_0} + \sqrt{\eta} \int_{t_0}^t e^{s\widetilde{A}_s} B dU_s. \right), \qquad \widetilde{A}_t := \begin{pmatrix} 3\frac{\log(t/t_0)}{t-t_0} I & H \\ -I & 0 \end{pmatrix}$$



(a)



(b)

Figure 4. MSGD vs SNAG (with constant momentum) dynamics for $\eta = 0.1$ and different $\lambda_d(H)$. (a) Dynamics at fixed $\mu = 0.2$. We observe that as predicted by the SME analysis, SNAG enjoys a faster linear convergence rate in the descent phase, as well as lower asymptotic fluctuations. The acceleration is indeed more pronounced for larger $\lambda_d(H)$. (b) When, μ for each case is chosen optimally for the descent (by maximizing the minimal real part of the eigenvalues of A_1, A_2 respectively), the dynamics becomes similar.

The eigenvalues of $\widetilde{A}_t$ are

$$\lambda(\widetilde{A}_t) = \left\{ \frac{1}{2} \left(3\frac{\log(t/t_0)}{t-t_0} \pm \sqrt{9[\frac{\log(t/t_0)}{t-t_0}]^2 - 4\lambda_i(H)} \right), \qquad i = 1, \ldots, d \right\}$$

Since there is no lower-bound on the minimal real part, the convergence is sub-linear. This is expected because the $\mathcal{O}(1/t)$ momentum schedule is suited for non-strongly-convex functions, whereas constant momentum is more appropriate for strong-convex functions [Nesterov(2013)]. Furthermore, we observe that since the real parts of all eigenvalues of $\widetilde{A}_t$ converge to 0 as $t \to \infty$, according to the analysis in Sec. 4.3, the asymptotic fluctuations due to noise should be large. Fig. 5 confirms these observations and further suggests that in the case of stochastic gradient methods, more careful momentum schedules must be derived in order to balance descent and fluctuations, e.g. using the optimal control framework presented in [Li et al.(2017)].

6 Conclusion

In this paper, we developed the mathematical foundations of the stochastic modified equations (SME) framework for analyzing stochastic gradient algorithms (SGAs). This approach is shown to be rigorous, flexible, and highly practical. By employing weak approximations, we provide a precise framework to bridge discrete stochastic gradient algorithms and continuous stochastic differential

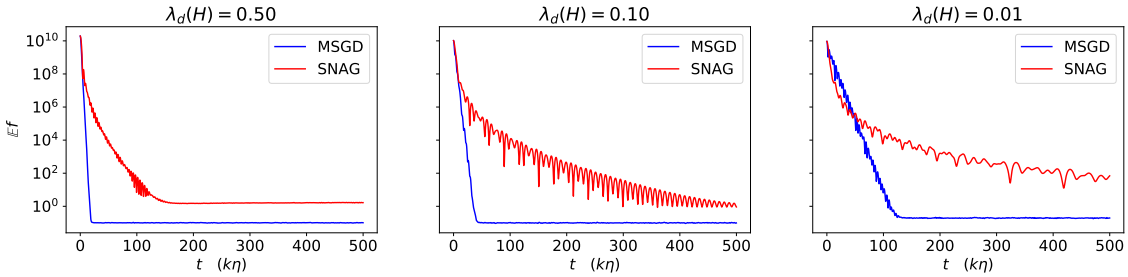


Figure 5. MSGD vs SNAG (with dynamic momentum according to Nesterov's choice (24)) dynamics for $\eta = 0.1$ and different $\lambda_d(H)$. We see that the convergence is indeed sub-linear, and moreover, the asymptotic fluctuations are large compared with MSGD, in which case μ here is picked to achieve optimal descent rate.

equations. Unlike strong approximations, which require path-wise accuracy, weak approximations allow for broader modeling flexibility, as demonstrated in Section 4.

Our primary result, outlined in Theorem 3, establishes a general framework for relating discrete-time algorithms to continuous-time SDEs, enabling the derivation of SMEs for various SGA variants. This flexibility was illustrated through the application of the SME framework to multiple algorithms, including momentum-based SGD, as discussed in Section 5.

Additionally, our numerical experiments in Section 5 highlighted the usefulness of the SME approach, particularly in analyzing the trade-off between gradient descent and stochastic fluctuations, and in understanding the performance differences between momentum-based and Nesterov-accelerated gradient methods.

In future work, we will extend the SME formalism to adaptive algorithms and further explore its practical applications in machine learning and optimization tasks.

References

[An et al.(2018)] Jing An, Jianfeng Lu, and Lexing Ying. Stochastic modified equations for the asynchronous stochastic gradient descent. *arXiv preprint arXiv:1805.08244*, 2018.

[Bach and Moulines(2013)] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[Betancourt et al.(2018)] Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

[Black and Scholes(1973)] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

[Daly(1963)] Bart J Daly. The stability properties of a coupled pair of non-linear partial difference equations. *Mathematics of Computation*, 17(84):346–360, 1963.

[Défossez and Bach(2015)] Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.

[Durrett(2010)] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[Evans(2010)] Lawrence C Evans. Partial differential equations. 2010.

[Feng et al.(2017)] Yuanyuan Feng, Lei Li, and Jian-Guo Liu. A note on semi-groups of stochastic gradient descent and online principal component analysis. *arXiv preprint arXiv:1712.06509*, 2017.

[Hirt(1968)] CW Hirt. Heuristic stability theory for finite-difference equations. *Journal of Computational Physics*, 2(4):339–355, 1968.

[Hu et al.(2017)] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

[Kloeden and Platen(2011)] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, New York, corrected edition, June 2011.

[Kushner and Yin(2003)] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[Kushner(1978)] Harold J Kushner. Rates of convergence for sequential monte carlo optimization methods. *SIAM Journal on Control and Optimization*, 16(1):150–168, 1978.

[Kushner and Shwartz(1984)] Harold J Kushner and Adam Shwartz. An invariant measure approach to the convergence of stochastic approximations with state dependent noise. *SIAM Journal on Control and Optimization*, 22(1):13–27, 1984.

[Kushner and Clark(2012)] Harold Joseph Kushner and Dean S Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

[Li et al.(2015)] Qianxiao Li, Cheng Tai, and Weinan E. Dynamics of stochastic gradient algorithms. arxiv preprint. *arXiv preprint arXiv:1511.06251v1*, 2015.

[Li et al.(2017)] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110, 2017.

[Ljung et al.(2012)] Lennart Ljung, Georg Ch Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012.

[Mandt et al.(2015)] Stephan Mandt, Matthew D Hoffman, and David M Blei. Continuous-time limit of stochastic gradient descent revisited. In *OPT workshop, NIPS*, 2015.

[Mandt et al.(2016)] Stephan Mandt, Matthew D Hoffman, and David M Blei. A variational analysis of stochastic gradient algorithms. *arXiv preprint arXiv:1602.02666*, 2016.

[Mandt et al.(2017)] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

[Milstein(1975)] Grigori N Milstein. Approximate integration of stochastic differential equations. *Theory of Probability & Its Applications*, 19(3):557–562, 1975.

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

[Moulines and Bach(2011)] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[Needell et al.(2014)] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized algorithm. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2014.

[Nesterov(2013)] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Nesterov(1983)] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

[Noh and Protter(1960)] WF Noh and MH Protter. Difference methods and the equations of hydrodynamics. Technical report, California. Univ., Livermore. Lawrence Radiation Lab., 1960.

[Oksendal(2013)] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[Shalev-Shwartz and Zhang(2014)] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pages 1–41, 2014.

[Shamir and Zhang(2013)] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

[Su et al.(2014)] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[Warming and Hyett(1974)] RF Warming and BJ Hyett. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of computational physics*, 14(2):159–179, 1974.

[Wibisono et al.(2016)] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.

[Xiao and Zhang(2014)] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.