Journal of Machine Learning Research 25 (2024) 1-55 Submitted 2/23; Revised 11/23; Published 1/24

A Comparison of Continuous-Time Approximations to Stochastic Gradient Descent

Stefan Ankirchner S.ANKIRCHNER@UNI-JENA.DE
Institute for Mathematics
Friedrich-Schiller-University Jena
07737 Jena, Germany

Stefan Perko STEFAN.PERKO@UNI-JENA.DE
Institute for Mathematics
Friedrich-Schiller-University Jena
07737 Jena, Germany

Editor: Stephan Mandt

Abstract

Applying a stochastic gradient descent (SGD) method for minimizing an objective gives rise to a discrete-time process of estimated parameter values. In order to better understand the dynamics of the estimated values, many authors have considered continuous-time approximations of SGD. We refine existing results on the weak error of first-order ODE and SDE approximations to SGD for non-infinitesimal learning rates. In particular, we explicitly compute the linear term in the error expansion of gradient flow and two of its stochastic counterparts, with respect to a discretization parameter h. In the example of linear regression, we demonstrate the general inferiority of the deterministic gradient flow approximation in comparison to the stochastic ones, for batch sizes which are not too large. Further, we demonstrate that for Gaussian features an SDE approximation with state-independent noise (CC) is preferred over using a state-dependent coefficient (NCC). The same comparison holds true for features of low kurtosis or large batch sizes. However, the relationship reverses for highly leptokurtic features or small batch sizes.

Keywords: stochastic gradient descent, gradient flow, stochastic differential equation, weak approximation, Talay-Tubaro expansion

1. Introduction

Consider a d-dimensional discrete-time stochastic process $\chi = (\chi_n)_{n\in\mathbb{N}_0}$ with dynamics

$$\chi_{n+1} = \chi_n - h\nabla R_{\gamma(n)}(\chi_n), \quad n \in \mathbb{N}_0, \tag{1}$$

where $(R_r)_r$ is a family of differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$, h is a positive real number, and $(\gamma(n))_{n\in\mathbb{N}_0}$ is an i.i.d. sequence of random variables. We interpret $(\chi^h_n)_{n\in\mathbb{N}_0}$ as the sequence of estimated parameters when applying a stochastic gradient descent (SGD) method for minimizing the function $\mathcal{R}(x) = \mathbb{E}[R_{\gamma(0)}(x)]$. The function $\mathcal{R}$ itself can be interpreted as *empirical risk* (that is training error) or *population risk*. We refer to h as the learning rate and $R_{\gamma(n)}$ as the risk due to the n-th data point or mini batch. In the following we simply call χ a SGD process.

To make the SGD process tractable with methods from mathematical analysis one frequently approximates the SGD dynamics with an ODE, usually referred to as gradient flow (GF), given by

$$dX_t^0 = b(X_t^0)\,dt, \quad X_0^0 = \chi_0, \tag{2}$$

where $b = -\nabla\mathcal{R}$. One can show that (2) is then a first-order approximation of SGD in the learning rate, that is for all $T > 0$ and nice test functions g we have

$$|\mathbb{E}g(\chi^h_{\lfloor T/h \rfloor}) - \mathbb{E}g(X_T^0)| = \mathcal{O}(h),$$

as $h \downarrow 0$.

GF dynamics are deterministic and hence ignore the randomness in SGD. Therefore, in recent years analytic approximations in terms of stochastic differential equations (SDEs) have become common. In particular, SDE approximations have been used to optimize hyperparameters (see Mandt et al., 2015, 2017; Li et al., 2017; Malladi et al., 2022), to analyze the long-term behavior of SGD processes (see Cao and Guo, 2020; Kunin et al., 2022; Wojtowytsch, 2024), to study the impact of normalization schemes (see Li et al., 2020)), to analyze the runtime until convergence (see Hu and Zhang, 2020), to study the transition between stationary points (see Yang et al., 2021; Zhou et al., 2020; Xie et al., 2020; Hu et al., 2017), to study the implicit bias and regularization properties of SGD (see Ali et al., 2020; Pesme et al., 2021; Li et al., 2022) and to study the effect of running SGD in parallel (see An et al., 2019; Boffi and Slotine, 2020).

Following Ali et al. (2020) we refer to solutions of SDEs approximating SGD as *stochastic gradient flow* (SGF). SGF dynamics are usually obtained by adding to the GF dynamics a diffusion term, typically driven by a Brownian motion W, and take the form

$$dX_t^h = b(X_t^h)\,dt + \sqrt{h}\sigma(X_t^h)\,dW_t. \tag{3}$$

Here, $\sigma(x) \in \mathbb{R}^{d\times d}$ denotes a symmetric positive semi-definite matrix. Two choices for σ are common: first, σ is constant, that is independent of the state (see Mandt et al., 2015); second, $\sigma(x)^2$ is equal to the covariance matrix of the sample gradient $\nabla\mathcal{R}_{\gamma(0)}(x)$ (see Li et al., 2017). We refer to a solution of (3) with constant σ as *constant covariance stochastic gradient flow* (CC-SGF), and a process with the second type of σ as *non-constant covariance stochastic gradient flow* (NCC-SGF).

However, without an additional modification of the *drift coefficient* b in Equation (3) the SGF dynamics are still merely a first-order approximation. In fact, by choosing any smooth σ of linear growth with bounded derivatives in (3), one obtains a weak approximation of order 1. Given that the order of approximation is not improved, does it make sense at all to add a diffusion term to the gradient flow dynamics? And if it does, how can one quantify the benefit?

To answer these questions, in this paper we expand the approximation errors of GF and (N)CC-SGF in h and compare their linear error terms, that is the constants in front of the linear term in the error expansion. It turns out that the linear error terms for GF, CC-SGF and NCC-SGF are generally all different. We can thus confirm a conjecture proposed in Feng et al. (2018, Remark 2.3.).

We characterize the linear error terms as integrals of functions applied to GF, hence our results bear similarities with the formulas of the leading weak error term when approximating SDEs with an Euler or Milstein scheme (see Talay and Tubaro, 1990). Indeed, Theorems

1, 2 and 3 can be seen as describing the leading term in the Talay-Tubaro expansion of the weak error. We remark, however, that the error estimate in the second and third theorem is given with respect to a *family* of SDEs, whereas the error considered by Talay and Tubaro (1990) refers to a *single* SDE.

Moreover, under a symmetry assumption and using the objective function as test function, we simplify these error terms further. This allows us to derive a bound on the linear error terms indicating that they tend to decrease as the curvature around the gradient flow trajectory increases.

Finally, we show that for linear regression models and constant learning rates, the linear error terms for the objective function can be calculated in closed form. A comparison then reveals that any of three continuous-time approximations can be the best, depending on the batch size. However, there is a notable caveat for the case of gradient flow being the best approximation. Note that the dynamics of learning a linear model using SGD with constant learning rate can be roughly separated into the initial *descent phase* and the final *fluctuation phase*, where SGD, due to the variance of the stochastic gradients, is mostly fluctuating around the global minimum. The batch size at which gradient flow becomes the best approximation increases as the duration of the fluctuation phase increases, relative to the time horizon. On the other hand, the approximation quality of the stochastic approximations is unaffected by the relative duration of the fluctuation phase.

1.1 Summary of Contributions

Below we provide a summary of the main contributions of this paper.

- We show that gradient flow (GF), stochastic gradient flow with constant covariance (CC-SGF) and stochastic gradient flow with non-constant covariance (NCC-SGF) are first-order approximations of SGD and related algorithms, with respect to the learning rate. In addition to previous works, we allow non-constant learning rates schedules which lead to time-inhomogeneous approximations. Furthermore, we derive an explicit expression for the linear error term in the error expansion with respect to the learning rate.

- For constant learning rates we express the three linear error terms using the first derivative ∇X^0 and second derivative $\nabla^2 X^0$ of the gradient flow with respect to its initial condition, as well as the first, second and third derivative of the objective function. We do this under the assumption that ∇X^0 is a symmetric matrix everywhere and that the test function g is the objective function to be minimized by SGD. As a consequence we obtain a natural bound on the linear error terms depending on the curvature along the gradient flow trajectory, the third derivative and the choice of diffusion coefficient.

- Using the linear error term expansion we study the example of linear regression with non-zero residuals, that is data noise, using population risk as test function. We show that there are two special batch sizes B^{Eq} and B^{GF}, such that for batch sizes $B < B^{\mathrm{Eq}}$ the NCC approximation is the best, followed by CC-SGF for $B^{\mathrm{Eq}} < B < B^{\mathrm{GF}}$ and GF for $B > B^{\mathrm{GF}}$. However, we also observe that B^{GF} increases with the duration of

the fluctation phase of SGD. On the other hand, B^{Eq} only depends on the kurtosis of the features.

1.2 Related Work

The idea to use stochastic differential equations for approximating SGD processes was first considered by Mandt et al. (2015) and Li et al. (2017, 2019). Mandt et al. (2015) heuristically use CC-SGF for approximating and analyzing the SGD process. Li et al. derived NCC-SGF (Li et al., 2017) and rigorously proved that it is a first-order approximation of SGD (Li et al., 2019). The approximation result is shown for constant learning rates and hence only for families of SDEs that are time-homogeneous. In contrast, our approximation results allow for time-dependent learning rates and give the linear error term explicitly.

Further results for the NCC-SGF approximation are derived by Lanconelli and Lauria (2022); Chen et al. (2020); Fontaine et al. (2021). Lanconelli and Lauria (2022) justifies the NCC-SGF dynamics with a general Markov chain convergence theorem. Theorem 3.5. by Chen et al. (2020) provides an estimate of the Wasserstein-1 distance between SGD processes and NCC-SGF. Fontaine et al. (2021) also considers NCC-SGF with time-dependent learning rates, assuming that the sequence of learning rates given by $\gamma(n+1)^{-\alpha}$ for some $\gamma \in (0,\infty)$ and $\alpha \in [0,1)$. Further, they provide an asymptotic estimate of the weak error as γ converges to zero (see Fontaine et al., 2021, Proposition 25). It is remarkable, that the same article also contains a strong approximation result (see Fontaine et al., 2021, Theorem 1) based on a coupling technique. In contrast to Fontaine et al. (2021), we provide explicit formulas for the linear error terms and we suppose less specific assumptions on the learning rate schedule u.

Moreover, the literature comprises articles considering weak approximations of order 2 for SGD processess (see Li et al., 2019; Feng et al., 2018, 2019; Gu et al., 2023).

Finally, we remark that our approximation results are asymptotic results, proving that (S)GF and SGD converge to each other as the learning rate converges to zero. The results do not provide any estimate of the actual error for fixed learning rates. That (S)GF may not be a good approximation of SGD if the learning rate is not sufficiently small is pointed out by Li et al. (2021).

2. General Results on Linear Error Terms

Let $d \in \mathbb{N}$ and $T > 0$. Given a subset D of Euclidean space, we write $g \in G(D)$ if $g : D \to \mathbb{R}$ has (at most) polynomial growth, that is there exists a constant $C > 0$ and $\kappa \in \mathbb{N}_0$, such that

$$|g(x)| \leq C(1 + |x|^{\kappa}) \tag{4}$$

for all x in the domain D of g. Typically, $D = \mathbb{R}^d$ or $D = [0,T] \times \mathbb{R}^d$. The infimum of all such C's for a given κ will be denoted by $\|g\|_{G_\kappa}$. We also sometimes write $g \in G_\kappa(D)$ if $\|g\|_{G_\kappa} < \infty$, especially for $\kappa = 1$. We write $g \in G^l(D)$ if $g \in C^l(D)$ and all its partial derivatives up to order l are in $G(D)$.

Now, let $(\Omega, \mathcal{F}_\Omega, \mathbb{P})$ be a complete probability space, Γ be a measurable space and $(\gamma(n))_{n\in\mathbb{N}_0}$ be a sequence of i.i.d. Γ-valued random variables. We can view $\gamma(n)$ as the data point or mini-batch chosen in the n-th iteration of stochastic gradient descent (SGD). Also

let $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ be a filtration on $(\Omega, \mathcal{F}_\Omega, \mathbb{P})$ independent of γ satisfying the usual conditions and let W be an $\mathbb{R}^d$-valued $\mathcal{F}$-Brownian motion.

Let $u : [0,T] \to [0,1]$ be a function.

Assumption (A1) *We have $u \in C^\infty$, such that u is constant or strictly decreasing.*

The function u is a learning rate schedule and represents the change of the learning rate over time. For all $h \in (0,1)$ the sequence of learning rates is given by $(hu_{nh})_{n\in\mathbb{N}_0}$. The parameter $h \in (0,1)$ acts as discretization parameter. If $\sup_{t\in[0,T]} u_t = 1$, then h can be interpreted as the *maximal* learning rate. Following Li et al. (2017) we have chosen to decompose the learning rate into h and u, because this makes the approximation results analogous to the autonomous (that is constant learning rate) case.

Recall that γ takes values in Γ. Let $H : \Gamma \times \mathbb{R}^d \to \mathbb{R}^d$. Now, given an initial value $x \in \mathbb{R}^d$, define (generalized) stochastic gradient descent by

$$\chi^h_{n+1} = \chi^h_n + hu_{nh}H_{\gamma(n)}(\chi^h_n), \quad \chi^h_0 = x. \tag{5}$$

Assumption (A2) *The function H satisfies $H \in G_1(\mathbb{R}^d)$ uniformly in $r \in \Gamma$, that is there exists a constant $C > 0$, such that*

$$|H_r(x)| \leq C(1 + |x|),$$

for all $r \in \Gamma$ and $x \in \mathbb{R}^d$.

Example 1 *The prototypical example to keep in mind is online SGD with replacement. Given a sequence of differentiable error functions $R_1, \dots, R_M : \mathbb{R}^d \to \mathbb{R}$, where M is the sample size of our data set, we set $H_{\gamma(n)}(x) := -\nabla R_{\gamma(n)}(x)$ and choose $\gamma(n)$ to be uniformly distributed on $\{1, \dots, M\}$.*

Finally, set

$$\bar{H} := \mathbb{E}H_{\gamma(0)} : \mathbb{R}^d \to \mathbb{R}^d,$$

and

$$\Sigma := \mathbb{E}[(H_{\gamma(0)} - \bar{H})^{\otimes 2}] : \mathbb{R}^d \to \mathbb{R}^{d\times d}.$$

Here $z^{\otimes 2} = zz^\dagger \in \mathbb{R}^{d\times d}$ for any $z \in \mathbb{R}^d$. By Assumption (A2) we have $\bar{H} \in G_1(\mathbb{R}^d)$.

Since Σ is symmetric and positive semi-definite, a unique matrix square root $\sqrt{\Sigma}$ exists.

Assumption (A3) *The functions $\bar{H}$ and $\sqrt{\Sigma}$ are Lipschitz continuous and in C^∞, such that all their partial derivatives are bounded.*

Next, we introduce three different (families of) differential equations, the solutions of which approximate the SGD iterations. Assumptions (A1) and (A3) guarantee that the coefficients of those equations are Lipschitz continuous. This assumption is standard in the literature on stochastic differential equations (SDEs) since it implies the existence and uniqueness of a solution. Assumptions (A1) and (A3) further imply the coefficients to be smooth with bounded partial derivatives. This property in turn implies smoothness of the SDE solutions with respect to the initial condition (see Theorem 27 in the Appendix or Kunita, 2004). Finally, Assumption (A2) yields linear growth and uniform integrability conditions on SGD and its increments (see Lemma 7 below).

2.1 Gradient Flow

Consider the ordinary differential equation

$$dX_t^0 = u_t \bar{H}(X_t^0)\, dt. \tag{6}$$

We will refer to equation (6) as (generalized) gradient flow, or GF for short.

The appearance of the learning rate schedule u in (6) may surprise. We are essentially considering the equidistant time grid $\{h, 2h, \dots, T\}$ in continuous-time. The discretization parameter h and the learning rate schedule u take on different roles, because the time grid is not affected by u. Let

$$\mathcal{H} := \{h \in (0,1) : T/h \in \mathbb{N}\} \tag{7}$$

be the set of acceptable learning rates and $g \in G^\infty(\mathbb{R}^d)$. For all $(t,x) \in [0,T] \times \mathbb{R}^d$ we define

$$v_t(x) = g(X_T^{0,t}(x)), \tag{8}$$

where $X^{0,t}(x)$ denotes the solution of (6) on $[t,T]$ with initial condition $X_t^t(x) = x$. Note that $X_T^{0,t}(x) = X_{T-t}^0(x)$. We write v_t^g if we want to emphasize the dependence of v on g. One can show that $v \in C^\infty([0,T] \times \mathbb{R}^d)$. Moreover, the partial derivatives of v with respect to time and space have polynomial growth in the space variable, uniformly in time. Hence, $v \in G^\infty([0,T] \times \mathbb{R}^d)$ in the sense that for every $k \in \mathbb{N}_0$ and multi-index[1] $\alpha \subseteq \{1, \dots, d\}$ there exist constants $C \in (0,\infty)$ and $\kappa \in \mathbb{N}_0$ such that

$$|\partial_t^k \partial_\alpha v_t(x)| \leq C(1 + |x|^\kappa), \tag{9}$$

for all $t \in [0,T]$ and $x \in \mathbb{R}^d$. Then, we define the function[2]

$$\varphi_t(x) = \frac{1}{2} u_t^2 \operatorname{tr}[\nabla^2 v_t(x) \bar{H}(x)^{\otimes 2}] + u_t \partial_t \nabla v_t(x)^\dagger \bar{H}(x) + \frac{1}{2} \partial_t^2 v_t(x), \tag{10}$$

with $(t,x) \in [0,T] \times \mathbb{R}^d$. Whenever we want to stress the dependence of φ on g we write φ^g.

Theorem 1 *Assume (A1), (A2) and (A3). Denote by X the solution of* (6) *with initial condition $X_0 = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$,*

$$\mathbb{E} g(\chi_{T/h}^h) - g(X_T^0) = h \int_0^T \varphi_t^g(X_t^0) + \frac{1}{2} u_t^2 \operatorname{tr}[\nabla^2 v_t^g(X_t^0) \Sigma(X_t^0)]\, dt + \mathcal{O}(h^2), \tag{11}$$

for all $h \in \mathcal{H}$, that is all discretization parameters h such that $\frac{T}{h}$ is an integer.

The parts of Assumption (A3) concerning $\sqrt{\Sigma}$ are superfluous for the proof of this theorem.

1. See the appendix before Theorem 27 for a definition of (unordered) multi-indices.
2. Here, ∇ denotes the gradient and ∇^2 the Hessian matrix with respect to x.

2.2 First-Order Stochastic Gradient Flow with Non-Constant Covariance

For all $h \in \mathcal{H} \cup \{0\}$ we consider the following family of stochastic differential equations, first introduced by Li et al. (2017),

$$dX_t^{\mathrm{NCC},h} = u_t \bar{H}(X_t^{\mathrm{NCC},h})\, dt + u_t \sqrt{h\Sigma(X_t^{\mathrm{NCC},h})}\, dW_t. \tag{12}$$

We refer to a process solving (12) as (generalized, first-order) *stochastic gradient flow with non-constant covariance* or NCC-SGF for short (in accordance with the terminology used by Ali et al., 2020). Notice that, as $h \downarrow 0$, the diffusion term in (12) vanishes and hence (12) becomes the ODE (6).

Theorem 2 *Assume (A1), (A2) and (A3). For all $h \in \mathcal{H}$ denote by X^h the solution of (12) with initial condition $X_0^h = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$ and $h \in \mathcal{H}$,*

$$\mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^{\mathrm{NCC},h}) = h \int_0^T \varphi_t^g(X_t^0)\, dt + \mathcal{O}(h^2), \tag{13}$$

where φ is defined in (10).

Note that the process X^0 is the same as gradient flow defined in (6).

2.3 First-Order Stochastic Gradient Flow with Constant Covariance

Finally, we consider an approximation to SGD with constant diffusion coefficient. Here, we have to make a choice on how to approximate Σ by a constant. Frequently one is interested in the behavior of SGD around a stationary point. In fact, suppose gradient flow converges to a, necessarily stationary, point $X_\infty^0 \in \mathbb{R}^d$. Then for every $h \in \mathcal{H} \cup \{0\}$ we consider the SDE

$$dX_t^{\mathrm{CC},h} = u_t \bar{H}(X_t^{\mathrm{CC},h})\, dt + u_t \sqrt{h\Sigma(X_\infty^0)}\, dW_t. \tag{14}$$

We refer to this approximation as (generalized, first-order) *stochastic gradient flow with constant covariance* or CC-SGF for short (again, in accordance with the terminology used by Ali et al., 2020). In the case $u = 1$ this is essentially the continuous-time approximation introduced by Mandt et al. (2015). Note that the diffusion coefficient may depend on the initial condition, since X_∞^0 may already depend on it.

Notice again that as $h \downarrow 0$ the diffusion term in (14) vanishes and hence (14) becomes the ODE (6).

Theorem 3 *Assume (A1), (A2), (A3) and that gradient flow converges to a stationary point X_∞^0. For all $h \in \mathcal{H}$ denote by $X^{\mathrm{CC},h}$ the solution of (14) with initial condition $X_0^{\mathrm{CC},h} = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$ and $h \in \mathcal{H}$,*

$$\begin{aligned}\mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^{\mathrm{CC},h}) =& h \int_0^T \varphi_t^g(X_t^0) + \frac{1}{2} u_t^2 \operatorname{tr}[\nabla^2 v_t^g(X_t^0)(\Sigma(X_t^0) - \Sigma(X_\infty^0))]\, dt \\ &+ \mathcal{O}(h^2),\end{aligned} \tag{15}$$

where v is defined in (8) and φ in (10).

2.4 Linear Error Terms for Minimization Problems

In specific settings we can give more explicit formulas for the linear error terms, building on top of Theorems 1, 2 and 3. In particular we consider SGD with $\bar{H} = -\nabla\mathcal{R}$, where $\mathcal{R}$ is some objective function to be minimized. For simplicity we consider only constant learning rates.

Note that (A3) implies $X^0 \in C^2([0,T]\times\mathbb{R}^d)$ if we consider gradient flow as a function of time and its initial condition. Therefore, we can consider the first and second derivative of gradient flow with respect to its initial condition. That is $\nabla X_t^0(x) \in \mathbb{R}^{d\times d}$ and $\nabla^2 X_t^0(x) \in \mathbb{R}^{d\times d\times d}$, where

$$\nabla X_t^0(x)_{i,j} = \partial_j X_t^0(x)_i, \quad \nabla^2 X_t^0(x)_{i,j,k} = \partial_{j,k} X_t^0(x)_i, \quad i,j,k \in \{1,\dots,d\},$$

for all $t \in [0,T]$ and $x \in \mathbb{R}^d$. More generally, if a function $f : \mathbb{R}^d \to \mathbb{R}^d$ is C^2, then we write $\nabla^2 f : \mathbb{R}^d \to \mathbb{R}^{d\times d\times d}$ where

$$(\nabla^2 f)_{i,j,k} = \partial_{j,k} f_i.$$

Similarly we define $\nabla^3 f = \nabla^2(\nabla f)$ for $f : \mathbb{R}^d \to \mathbb{R}$. For $k \in \mathbb{N}_0$ we write

$$d^{\times k} = \underbrace{d \times \cdots \times d}_{k \text{ times}}.$$

Given $k, l \in \mathbb{N}_0$ as well as tensors $A \in \mathbb{R}^{d^{\times(k+l)}}$ and $B \in \mathbb{R}^{d^{\times l}}$, we define $\langle A, B\rangle \in \mathbb{R}^{d^{\times k}}$ by summing over the common indices, that is

$$\langle A, B\rangle_{i_1,\dots,i_k} := \sum_{j_1,\dots,j_l} A_{i_1,\dots,i_k,j_1,\dots,j_l} B_{j_1,\dots,j_l}.$$

In particular, given vectors $u, v \in \mathbb{R}^d$ and matrices $A, B \in \mathbb{R}^{d\times d}$ we have

$$\langle u, v\rangle = u^\dagger v, \quad \langle A, B\rangle = \operatorname{tr}(A^\dagger B) \in \mathbb{R}.$$

The quantity $\langle A, B\rangle$ is also known as the Frobenius inner product of A and B. The inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^{d^{\times k}}$ induces the Frobenius norm, given by $\|A\|_F = \sqrt{\langle A, A\rangle}$ for all $A \in \mathbb{R}^{d^{\times k}}$.

Given a continuous-time (stochastic) approximation $Y = (Y_t^h)_{t\in[0,T], h\in\mathcal{H}}$ of SGD we define the linear error term (with respect to $\mathcal{R}$) by

$$\mathrm{LE}(Y) := \lim_{h\downarrow 0} \frac{\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(Y_T^h)}{h},$$

where the limit is taken in $\mathcal{H}$.

Theorem 4 *Assume $u = 1$, (A2) and (A3). Further, assume we are given $\mathcal{R} \in C^\infty(\mathbb{R}^d)$, such that $\bar{H} = -\nabla\mathcal{R}$ and ∇X_t^0 is a symmetric matrix for all $t \geq 0$ and initial values. Let $x \in \mathbb{R}^d$, $D \in \{0, \Sigma(X_\infty^0), \Sigma\}$ and[3] consider the solution X to the family of stochastic differential equations*

$$dX_t^h = -\nabla\mathcal{R}(X_t^h)\,dt + \sqrt{hD(X_t^h)}\,dW_t, \quad X_0^h = x, \quad t \in [0,T], h \in (0,1).$$

3. In the case $D = \Sigma(X_\infty^0)$ we implicitly assume that the limit X_∞^0 exists.

Then we have

$$\mathrm{LE}(X) = \frac{1}{2}\langle \nabla^2 \mathcal{R}(X_T^0), \alpha_T^D - T\nabla\mathcal{R}(X_T^0)^{\otimes 2}\rangle + \frac{1}{2}\langle \nabla\mathcal{R}(X_T^0), \beta_T^D\rangle, \tag{16}$$

where

$$\alpha_T^D = \int_0^T \nabla X_T^{0,t}(X_t^0)(\Sigma(X_t^0) - D(X_t^0))\nabla X_T^{0,t}(X_t^0)\, dt \in \mathbb{R}^{d\times d},$$
$$\beta_T^D = \int_0^T \langle \nabla^2 X_T^{0,t}(X_t^0), \nabla\mathcal{R}(X_t^0)^{\otimes 2} + \Sigma(X_t^0) - D(X_t^0)\rangle\, dt \in \mathbb{R}^d.$$

Moreover,

$$\begin{aligned} 2|\,\mathrm{LE}(X)| \leq & dTM_T^2|\nabla\mathcal{R}(x)|^2(\|\nabla^2\mathcal{R}(X_T^0)\|_F + d^{3/2}|\nabla\mathcal{R}(x)|\zeta_T) \\ & + d\xi_T^D(\|\nabla^2\mathcal{R}(X_T^0)\|_F + \sqrt{d}|\nabla\mathcal{R}(x)|\zeta_T), \end{aligned} \tag{17}$$

where

$$M_t = \exp\left(-\int_0^t \lambda_{\min}(\nabla^2\mathcal{R}(X_s^0))\, ds\right), \quad \zeta_t = \int_0^t M_s\|\nabla^3\mathcal{R}(X_s^0)\|_F\, ds,$$
$$\xi_t^D = M_t^2\int_0^t \frac{\|\Sigma(X_s^0) - D(X_s^0)\|_F}{M_s^2}\, ds, \quad t\in[0,T].$$

Note that in the case of the NCC approximation we have $\xi_T^D = \xi_T^\Sigma = 0$. The term M_T goes to 0, as the curvature, that is the smallest eigenvalue of the Hessian matrix, along the gradient flow trajectory becomes large. Thus, as long as ζ_T does not grow too rapidly as curvature increases, the linear error term of NCC-SGF vanishes as curvature becomes large. This observation is analogous to a result by Elkabetz and Cohen (2021) suggesting that "the "more convex" the objective function is around the gradient flow trajectory, the better the match between gradient flow and gradient descent is guaranteed to be." However, they only consider deterministic gradient descent and the gradient flow approximation. Curiously, the third derivative of the objective function (and by extension ζ_T) also does not seem to play a role in their theory.

We can also make rough statements to compare the three approximations in general. Note that either the first or second summand in (16) dominates the other in size. In any case, by considering $D = \Sigma(X_\infty^0)$ we can see that the absolute linear error term for the CC-SGF approximation becomes large as $F(x) = \Sigma(x) - \Sigma(X_\infty^0)$, that is the noise outside of the stationary point, becomes large. Therefore, for large F the NCC approximation is better than CC-SGF. By considering $D = 0$ and writing $\Sigma(x) = F(x) + \Sigma(X_\infty^0)$ we also see that NCC-SGF is better than gradient flow for large F. Moreover, for large noise around the stationary point that gradient flow approaches, that is large $\Sigma(X_\infty^0)$, Theorem 4 further indicates that the stochastic approximations are preferable to gradient flow.

In the next section we will use Theorem 4 to analyze and compare the three continuous-time approximation more precisely in the setting of linear regression.

3. A Comparison of Continuous-Time Approximations to SGD for Linear Regression

In this section we compare gradient flow and the two stochastic gradient flow approximations (NCC and CC) in the setting of linear regression using mini-batch SGD. For simplicity we only consider constant learning rates in this section, that is $u = 1$.

Firstly, we provide a theoretical comparison using Theorems 1, 2 and 3 (see Theorem 6). We will see that the comparison highly depends on the batch size and on the kurtosis of the features (also called independent variables). Secondly, we substantiate the theoretical findings using a numerical example.

In a fairly general, parametric, statistical learning setting we are given an unknown measure ν, called *population*, on a measurable space $\mathcal{Z}$, a set of parameters $\Theta \subseteq \mathbb{R}^d$ and a family of risk functions $(R_z(\theta))_{\theta\in\Theta, z\in\mathcal{Z}}$. The general goal of statistical learning is then to minimize over Θ the *population risk*, that is the mean risk of the data under the measure ν

$$\mathcal{R}(\theta) := \mathbb{E}_{z\sim\nu}[R_z(\theta)].$$

Accordingly, we focus on comparing the weak error of the continuous-time approximations of SGD for the population risk function $\mathcal{R}$ associated with a linear regression task.

In terms of our interpretation and in our examples we focus on this "population setting", where we are essentially performing SGD without replacement for an infinite sequence[4] of i.i.d. data.

3.1 The Statistical Learning Setting

In this subsection we introduce the statistical learning setting in the case of linear regression. Suppose we are given an $\mathbb{R}^d$-valued random variable $\boldsymbol{x}$ and an $\mathbb{R}$-valued random variable $\boldsymbol{\varepsilon}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ are independent, $\mathbb{E}\boldsymbol{\varepsilon} = 0, \sigma_\varepsilon^2 := \mathbb{E}\boldsymbol{\varepsilon}^2 < \infty$, the covariance matrix κ of $\boldsymbol{x}$ is positive definite, and $\boldsymbol{x}$ has finite joint fourth moments

$$\mathbb{E}|\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{x}_k\boldsymbol{x}_l| < \infty, \quad i,j,k,l \in \{1,\dots,d\}.$$

Let $\theta^* \in \mathbb{R}^d$. We define the $\mathbb{R}$-valued random variable $\boldsymbol{y}$ by

$$\boldsymbol{y} = \langle \theta^*, \boldsymbol{x}\rangle + \boldsymbol{\varepsilon}.$$

Denote the distribution of $(\boldsymbol{x}, \boldsymbol{y})$ by ν. We call ν the *population*. We consider data drawn from ν, which follows a linear model. The population is considered unknown to us.

Note that in this section we follow the convention from statistics and denote the features (explanatory variables) by x (or $\boldsymbol{x}$ if they are random). The initial condition of SGD and its approximations is here denoted by θ (and not by x in contrast to the previous sections).

Let ℓ be the *square loss*, given by $\ell(y, y') = \frac{1}{2}(y - y')^2$. The goal is to fit the data drawn from ν using a linear predictor $\theta \mapsto \langle \theta, x\rangle$. Thus, for any data point $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ we consider the squared risk

$$R_{x,y}(\theta) = \ell(\langle\theta, x\rangle, y) = \frac{1}{2}(\langle\theta, x\rangle - y)^2.$$

4. Only finitely many samples are used, since we are given fixed time horizon.

We define the *population risk* by

$$\mathcal{R}(\theta) := \mathbb{E}[R_{\boldsymbol{x},\boldsymbol{y}}(\theta)].$$

We stress that the bold letters $\boldsymbol{x}, \boldsymbol{y}$ denote random variables, while x, y represent realizations. The minimum of $\mathcal{R}$, that is the best possible fit, is given by the population parameter θ^*. Fix a batch size $B \in \mathbb{N}$. We can determine an estimate of θ^* using mini-batch stochastic gradient descent

$$\begin{aligned}\chi^h_{n+1} =& \chi^h_n - \frac{h}{B}\sum_{k=0}^{B-1} \nabla_\theta R_{\boldsymbol{x}_{k+Bn},\boldsymbol{y}_{k+Bn}}(\chi^h_n) \\ =& \chi^h_n - \frac{h}{B}\sum_{k=0}^{B-1} (\langle \chi^h_n, \boldsymbol{x}_{k+Bn}\rangle - \boldsymbol{y}_{k+Bn})\boldsymbol{x}_{k+Bn}, \end{aligned} \tag{18}$$

where $(\boldsymbol{x}_n, \boldsymbol{y}_n)_{n\in\mathbb{N}_0}$ is an i.i.d.sequence with $(\boldsymbol{x}_0, \boldsymbol{y}_0) \sim \nu$. We calculate

$$\mathcal{R}(\theta) = \frac{1}{2}\mathbb{E}[(\langle \theta - \theta^*, \boldsymbol{x}\rangle - \boldsymbol{\varepsilon})^2] = \frac{1}{2}\langle \kappa, (\theta-\theta^*)^{\otimes 2}\rangle + \frac{\sigma_\varepsilon^2}{2}.$$

Hence, $\mathcal{R}$ is a quadratic form with minimum at θ^*. Now, consider $D \in \{0, \Sigma(\theta^*), \Sigma\}$ and

$$dX^h_t = -\nabla\mathcal{R}(X^h_t) + \sqrt{hD(X^h_t)}\, dW_t, \quad t \in [0,T], h \in (0,1).$$

One can show (see Section 6.1) the linear error term is given by

$$\mathrm{LE}(X) = -\frac{1}{2}T\langle \kappa^3 e^{-2T\kappa}, (\theta-\theta^*)^{\otimes 2}\rangle + \frac{1}{2}\langle \kappa, \alpha_T^D\rangle,$$

where

$$\alpha_T^D = \int_0^T e^{-(T-t)\kappa}(\Sigma(X^0_t) - D(X^0_t))e^{-(T-t)\kappa}\, dt.$$

To calculate α_T we will start with the covariance matrix of the gradient noise at batch size 1, which is given by

$$S(\theta) := \mathrm{Cov}[\nabla_\theta R_{\boldsymbol{x},\boldsymbol{y}}(\theta)] = \langle \mu_x^4 - \kappa^{\otimes 2}, (\theta-\theta^*)^{\otimes 2}\rangle + \sigma_\varepsilon^2\kappa$$

where $\mu_x^4 \in \mathbb{R}^{d\times d\times d\times d}$ with

$$(\mu_x^4)_{i,j,k,l} = \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{x}_k\boldsymbol{x}_l], \quad i,j,k,l \in \{1,\dots,d\}.$$

Note that the covariance matrix of the gradient noise is given by $\Sigma(\theta) = \frac{1}{B}S(\theta)$.

In the next subsection we add a natural assumption on S that allows for computation of explicit linear error terms not involving any integrals.

3.2 Theoretical Comparison of the Linear Error Terms

In this subsection we will give a comparison of the linear error terms of the three continuous-time approximations of SGD for linear regression. We will see that the comparison depends on the batch size B and the kurtosis of the features. Before we start we narrow down the assumptions further to two particular settings in order to simplify the covariance matrix S.

Example 2 *We study in detail the following two specific settings.*

(a) We assume that the features are centered Gaussian, that is $\boldsymbol{x} \sim \mathcal{N}(0, \kappa)$. Then we can simplify the covariance matrix of the gradient noise for batch size 1 to

$$S(\theta) = 2\kappa(\theta - \theta^*)^{\otimes 2}\kappa + \sigma_\varepsilon^2 \kappa.$$

(b) We assume that $d = 1$, but not that $\boldsymbol{x}$ is Gaussian. Then, we can write

$$S(\theta) = \kappa^2(\operatorname{Kurt} \boldsymbol{x} - 1)(\theta - \theta^*)^2 + \kappa\sigma_\varepsilon^2,$$

where $\operatorname{Kurt} \boldsymbol{x} := \mathbb{E}[\boldsymbol{x}^4]/\kappa^2$ *is the* kurtosis *of $\boldsymbol{x}$ (see Section A in the appendix for more information about kurtosis).*

From now on, assume that there exists a constant $B^{\mathrm{Eq}} > 0$, such that

$$S(\theta) = 2B^{\mathrm{Eq}}\kappa(\theta - \theta^*)^{\otimes 2}\kappa + \sigma_\varepsilon^2\kappa, \quad \theta \in \mathbb{R}^d.$$

In particular, in Example 2 (a) we have $B^{\mathrm{Eq}} = 1$ and for (b) we have $B^{\mathrm{Eq}} = \frac{1}{2}(\operatorname{Kurt} \boldsymbol{x} - 1)$. Proposition 5 below implies that if $B^{\mathrm{Eq}} \in \mathbb{N}$, then it is the batch size B where the NCC and CC approximation have the same error, up to flipping the sign.

Now, the three continuous-time approximations (6), (12) and (14) take the form

$$\begin{aligned}
dX_t^0 &= -\kappa(X_t^0 - \theta^*)\,dt \\
dX_t^{\mathrm{NCC},h} &= -\kappa(X_t^{\mathrm{NCC},h} - \theta^*)\,dt + \sqrt{\frac{h}{B}}\sqrt{2B^{\mathrm{Eq}}\kappa(X_t^{\mathrm{NCC},h} - \theta^*)^{\otimes 2}\kappa + \sigma_\varepsilon^2\kappa}\,dW_t \\
dX_t^{\mathrm{CC},h} &= -\kappa(X_t^{\mathrm{CC},h} - \theta^*)\,dt + \sqrt{\frac{h}{B}\sigma_\varepsilon^2\kappa}\,dW_t.
\end{aligned} \tag{19}$$

Note that the process with constant covariance dynamics (19) is an Ornstein-Uhlenbeck process. Using (42) we can derive the following expressions for the linear error terms of the three continuous-time approximations of SGD.

Proposition 5 *Suppose $\chi_0^h = X_0^0 = X_0^{\mathrm{NCC},h} = X_0^{\mathrm{CC},h} = \theta \in \mathbb{R}^d$ for all $h \in \mathcal{H}$. Then, we have*

$$\begin{aligned}
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^h) &= -\frac{h}{2}T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle + \mathcal{O}(h^2), \\
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^{\mathrm{CC},h}) &= h\left(\frac{B^{\mathrm{Eq}}}{B} - \frac{1}{2}\right)T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle + \mathcal{O}(h^2), \\
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^0) &= h\left(\frac{B^{\mathrm{Eq}}}{B} - \frac{1}{2}\right)T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle \\
&\quad + \frac{h}{4B}\sigma_\varepsilon^2\langle \kappa, 1_{d\times d} - e^{-2\kappa T}\rangle + \mathcal{O}(h^2).
\end{aligned} \tag{20}$$

as $h \downarrow 0$, with T/h an integer.

We introduce some additional notation to succinctly state the following theorem. Given two continuous-time approximations Y, Z we write $Y \preceq Z$ if $|\operatorname{LE}(Y)| \geq |\operatorname{LE}(Z)|$, that is if the approximation of SGD with Y has (in absolute terms) a greater linear error term than the one using Z. More briefly it means that Z is not worse than Y. Evidently $\preceq$ is a reflexive and transitive relation. We write $Y \asymp Z$ if $Y \preceq Z$ and $Z \preceq Y$, that is if Y and Z are equally good approximations. Further, we write $Y \prec Z$ if $Y \preceq Z$ and $Z \not\preceq Y$, that is if Z is strictly a better approximation than Y.

Theorem 6 *Suppose $B^{\mathrm{Eq}} > 0$ and we are given an initial value $\theta \neq \theta^*$. Define*

$$B^{\mathrm{GF}} = 2B^{\mathrm{Eq}} + \frac{\sigma_\varepsilon^2 \langle \kappa, 1 - e^{-2T\kappa} \rangle}{4T \langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2} \rangle}.$$

Then, we have the following

(i) $X^0 \prec X^{\mathrm{CC}} \prec X^{\mathrm{NCC}}$, *if* $B < B^{\mathrm{Eq}}$,

(ii) $X^0 \prec X^{\mathrm{CC}} \asymp X^{\mathrm{NCC}}$, *if* $B = B^{\mathrm{Eq}}$,

(iii) $X^0 \prec X^{\mathrm{NCC}} \prec X^{\mathrm{CC}}$, *if* $B^{\mathrm{Eq}} < B < B^{\mathrm{GF}} - B^{\mathrm{Eq}}$,

(iv) $X^{\mathrm{NCC}} \prec X^0 \prec X^{\mathrm{CC}}$, *if* $B^{\mathrm{GF}} - B^{\mathrm{Eq}} < B < B^{\mathrm{GF}}$,

(v) $X^{\mathrm{NCC}} \prec X^{\mathrm{CC}} \prec X^0$, *if* $B > B^{\mathrm{GF}}$,

(vi) $\operatorname{LE}(X^{\mathrm{CC}}) = 0$, *if* $B = 2B^{\mathrm{Eq}}$.

In other words, for small batch sizes the best approximation is NCC-SGF, followed by CC-SGF and then gradient flow. If we increase the batch size, then NCC and CC switch places. After that NCC and GF switch places. Finally, for large batch sizes GF becomes the best approximation. Somewhere in between CC is not only the best approximation among the three, but also has a linear error of 0.

Even though the gradient flow approximation can be the best approximation for large batch sizes, the lower bound B^{GF} for this to occur diverges to ∞ as

$$T \to \infty, \text{ or } \sigma_\varepsilon \to \infty, \text{ or } \kappa \to \infty \text{ (for } d = 1\text{), or } \theta - \theta^* \to 0 \text{ (for } d = 1\text{).} \tag{21}$$

In fact, one can summarize (21) by saying $\tau \to \infty$, where τ is the time that SGD spends fluctuating around the global minimum θ^*. Therefore, for large τ the SGF approximations are preferable to gradient flow, for all reasonably large batch sizes.

When it comes to deciding between NCC and CC-SGF, the important quantity is B^{Eq}. This quantity only depends on the distribution of $\boldsymbol{x}$ and *not* on $T, \kappa, \sigma_\varepsilon$ or $\theta - \theta^*$. For $\boldsymbol{x}$ Gaussian we have $B^{\mathrm{Eq}} = 1$, so the CC-SGF approximation is, perhaps surprisingly, almost always preferred over the NCC approximation. We also consider the case where $d = 1$ and $B^{\mathrm{Eq}} = \frac{1}{2}(\operatorname{Kurt} x - 1)$. In this case we observe for batch sizes that are small, relative to the kurtosis of the features $\boldsymbol{x}$, the NCC approximation can still be the best one (see also Section A in the appendix for more information on kurtosis).

Overall, one can also say that for highly leptokurtic features, the NCC approximation is the best across a large range of batch sizes. On the other hand, for lower kurtosis the CC approximation is best.

Figure 1 below provides a visual comparison of the three approximations in terms of kurtosis and batch size in two specific examples.

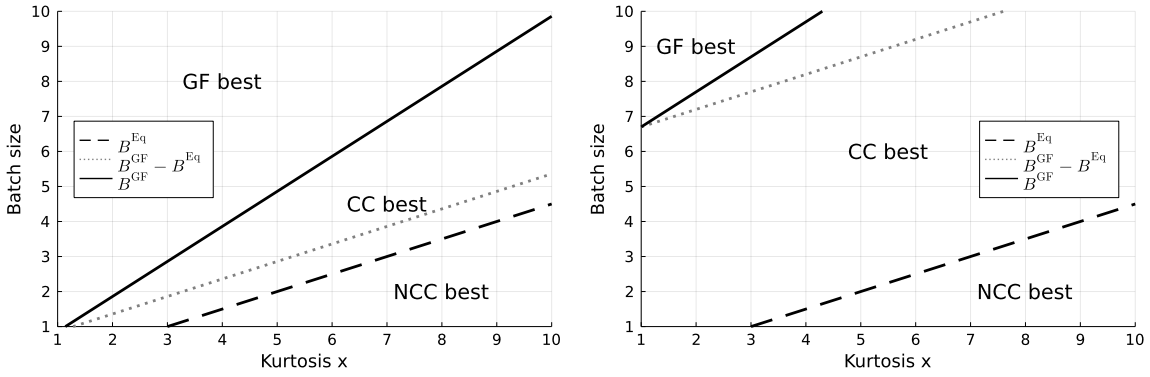

Figure 1: The best continuous-time approximation of SGD for linear regression in dimension 1 in terms of the kurtosis of the features and the batch size. Here $\kappa = 1$, $(\theta - \theta^*)^2 = 1$ and $T = 0.5$ ($T = 2.0$) in the left (right) plot. In the lower part of the middle region, where CC-SGF is the best approximation, gradient flow is worse than NCC-SGF. In the upper part of the middle region, gradient flow is better than NCC-SGF.

3.2.1 The Case of Batch Size 1

Here, we specifically study the case $B = 1$. Firstly, we have

$$X^0 \prec X^{\mathrm{CC}} \asymp X^{\mathrm{NCC}}, \text{ if } \boldsymbol{x} \text{ is Gaussian.}$$

Secondly, if $d = 1$ and so $B^{\mathrm{Eq}} = \frac{1}{2}(\mathrm{Kurt}\,\boldsymbol{x} - 1)$, then

(i) $X^0 \prec X^{\mathrm{CC}} \prec X^{\mathrm{NCC}}$, if $\mathrm{Kurt}\,\boldsymbol{x} > 3$,

(ii) $X^0 \prec X^{\mathrm{CC}} \asymp X^{\mathrm{NCC}}$, if $\mathrm{Kurt}\,\boldsymbol{x} = 3$,

(iii) $X^{\mathrm{NCC}} \prec X^{\mathrm{CC}}$, if $\mathrm{Kurt}\,\boldsymbol{x} \in (1, 3)$,

(iv) $X^{\mathrm{NCC}} = X^{\mathrm{CC}}$, if $\mathrm{Kurt}\,\boldsymbol{x} = 1$,

(v) $\mathrm{LE}(X^{\mathrm{CC}}) = 0$, if $\mathrm{Kurt}\,\boldsymbol{x} = 2$.

Note that distributions with kurtosis < 3 / $= 3$ / > 3 are also called platykurtic / mesokurtic / leptokurtic (see also Section A)

Gradient flow is always the worst approximation for $\mathrm{Kurt}\,x \geq 3$. Assume we are in the platykurtic setting $\mathrm{Kurt}\,\boldsymbol{x} \in (1, 3)$. Then gradient flow is the worst / second-best / best approximation if

$$1 < B^{\mathrm{GF}} - B^{\mathrm{Eq}} \quad / \quad B^{\mathrm{GF}} - B^{\mathrm{Eq}} < 1 < B^{\mathrm{GF}} \quad / \quad B^{\mathrm{GF}} < 1.$$

3.3 A Numerical Example

In this subsection we present results from a numerical experiment confirming the theoretical results presented in Theorem 6. We also compare the three approximations to yet another

continuous-time approximation to SGD, which we call *second-order stochastic gradient flow*, or SGF2 for short. The corresponding family of stochastic differential equations is given by

$$\begin{aligned} dX_t^{2,h} &= -\mathcal{R}'(X_t^{2,h}) - \frac{h}{2}\mathcal{R}''(X_t^{2,h})\mathcal{R}'(X_t^{2,h})\,dt + \sqrt{h\Sigma(X_t^{2,h})}\,dW_t \\ &= -\kappa\left(1_{d\times d} + \frac{h}{2}\kappa\right)(X_t^h - \theta^*)\,dt + \sqrt{\frac{h}{B}}\sqrt{2B^{\mathrm{Eq}}\kappa(X_t^{2,h}-\theta^*)^{\otimes 2}\kappa + \sigma_{\tilde{\varepsilon}}^2\kappa}\,dW_t \end{aligned} \tag{22}$$

with $X_0^{2,h} = \chi_0$. Then the following holds: for every $T > 0$ and $g \in G^\infty(\mathbb{R})$ there exists a $C > 0$, such that (see Li et al., 2019)

$$|\mathbb{E}g(\chi^h_{\lfloor T/h \rfloor}) - \mathbb{E}g(X_T^{2,h})| \leq Ch^2. \tag{23}$$

In other words, the linear error term is 0 (regardless of whether $g = \mathcal{R}$ or not). In this sense SGF2 is the best approximation we have seen so far. To achieve this improvement we use the same diffusion coefficient Σ as NCC-SGF, while making the drift coefficient more complicated.

For the remainder of this section we exclusively work in setting (b) from Example 2.

3.3.1 Experimental Setup

We consider using SGD for fitting the particular one-dimensional linear model

$$\boldsymbol{y} = -\boldsymbol{x} + \boldsymbol{\varepsilon} \tag{24}$$

with $\boldsymbol{x}, \boldsymbol{\varepsilon}$ independent, centered and of variance 1, where $\boldsymbol{\varepsilon}$ is Gaussian. Note that in this case we have $\theta^* = -1$. We compare the weak errors of the population risk $\mathcal{R}$ for different continuous-time approximations of SGD. Here we use time horizons $T = 0.5$ and $T = 2.0$, varying distributions of $\boldsymbol{x}$ and initial values θ. We use a Monte Carlo approximation to estimate $\mathbb{E}\mathcal{R}(\chi^h_{T/h})$, that is

$$\mathbb{E}\mathcal{R}(\chi^h_{T/h}) \approx \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\mathcal{R}(\hat{\chi}^{i,h}_{T/h})$$

where $\hat{\chi}^1, \ldots, \hat{\chi}^M$ are independent copies of χ. More precisely, to compute one copy $\hat{\chi}^i$ we draw BT/h i.i.d. samples from the data-generating model (24) and then perform SGD for T/h steps using a batch of B samples in each step, never using any sample twice. Thus, every copy of $\hat{\chi}$ uses a different (pseudo-) data set. For the experiments we have chosen M large enough (between 10^8 and $2 \cdot 10^9$) so that the variance of the Monte Carlo estimator is negligible compared to the weak error. Moreover, to reduce the computational burden significantly, we determine $\mathbb{E}\mathcal{R}^e(Y_T^h)$ for $Y = X^0, X^{\mathrm{NCC}}h, X^{\mathrm{CC},h}, X^{2,h}$ using explicit formulas, which can be derived in this example (see Proposition 25 in Section 6.3). We consider the learning rates $h = 0.5, 0.1, 0.05, 0.01, 0.005, 0.001$. Notice that T/h is an integer in each case, where $T \in \{0.5, 2.0\}$. Plotted is the dependence of the weak error

$$\frac{1}{\kappa}|\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(Y_T^h)|,$$

divided by κ (!), on the learning rate h.

3.3.2 Results

In the following ν_x denotes any distribution with expectation m, such that $\boldsymbol{x} + m \sim \nu_x$. That is $\boldsymbol{x}$ has distribution ν_x, but shifted to have expectation zero. Figure 2 depicts the weak error's dependence on the learning rate in the following settings:

Nr	T	θ	ν_x	κ	Kurt $\boldsymbol{x}$	B	B^{Eq}	$B^{\mathrm{GF}} - B^{\mathrm{Eq}}$	B^{GF}
(1)	0.5	0	Exp(0.1)	10	9	1	4	114.127	118.127
(2)	0.5	0	$\mathcal{N}(0,1)$	1	3	1	1	1.85914	2.85914
(3)	2.0	0	$\mathcal{N}(0,1)$	1	3	4	1	7.69977	8.69977
(4)	0.5	0	Exp(1)	1	9	8	4	4.85914	8.85914
(5)	0.5	0	$\mathcal{N}(0,1)$	1	3	4	1	1.85914	2.85914
(6)	0.5	-0.9	$\mathcal{N}(0,1)$	1	3	2	1	86.9141	87.9141

Aside from minor deviations stemming from the Monte Carlo estimation, the empirical results in Figure 2 confirm the theoretical results in the last subsection. In particular, we observe:

(i) The experimental settings (1)—(5) correspond exactly to the settings (i)—(v) in Theorem 6. Note that instead of merely varying the batch size B we also varied B^{Eq} and B^{GF} by choosing different T and distributions of $\boldsymbol{x}$.

(ii) As indicated by Proposition 5, the experimental setting (6) shows that for $\theta \approx \theta^*$ and only moderately small learning rates there is little difference between the NCC- and the CC-SGF approximations, while gradient flow is lagging behind by neglecting to model the variance of the residuals σ_ε^2.

(iii) For $B = B^{\mathrm{Eq}}$, NCC- and CC-SGF are equally good (setting (2)).

(iv) For $B = 2B^{\mathrm{Eq}}$ the CC-SGF approximation is of second order[5] (settings (4) and (6)).

(v) The SGF2 approximation is always best, irrespective of batch size.

We remark that the theoretical rates of convergence are difficult to observe without using a high number of Monte Carlo samples. Moreover, note that in the experiments we always plotted the weak error while Theorem 6 only applies to the linear error term. The results indicate that the higher order error terms have negligible impact on the total error.

3.4 Generalizations

One may wonder how much the results from this section generalize beyond the particular setting of linear regression using SGD without replacement. For example, a more commonly studied example is that of SGD *with* replacement, where the randomness stems from the sampling procedure and not from the underlying population measure. Then the objective function $\mathcal{R}$ is the empirical risk, also called training error. In this case the covariance matrix is still, in some sense, given by a quadratic form (of matrices). Proposition 5 still holds true,

5. More precisely, the approximation is of order 2 for the chosen test function $\mathcal{R}$. This is a weaker property than (23).

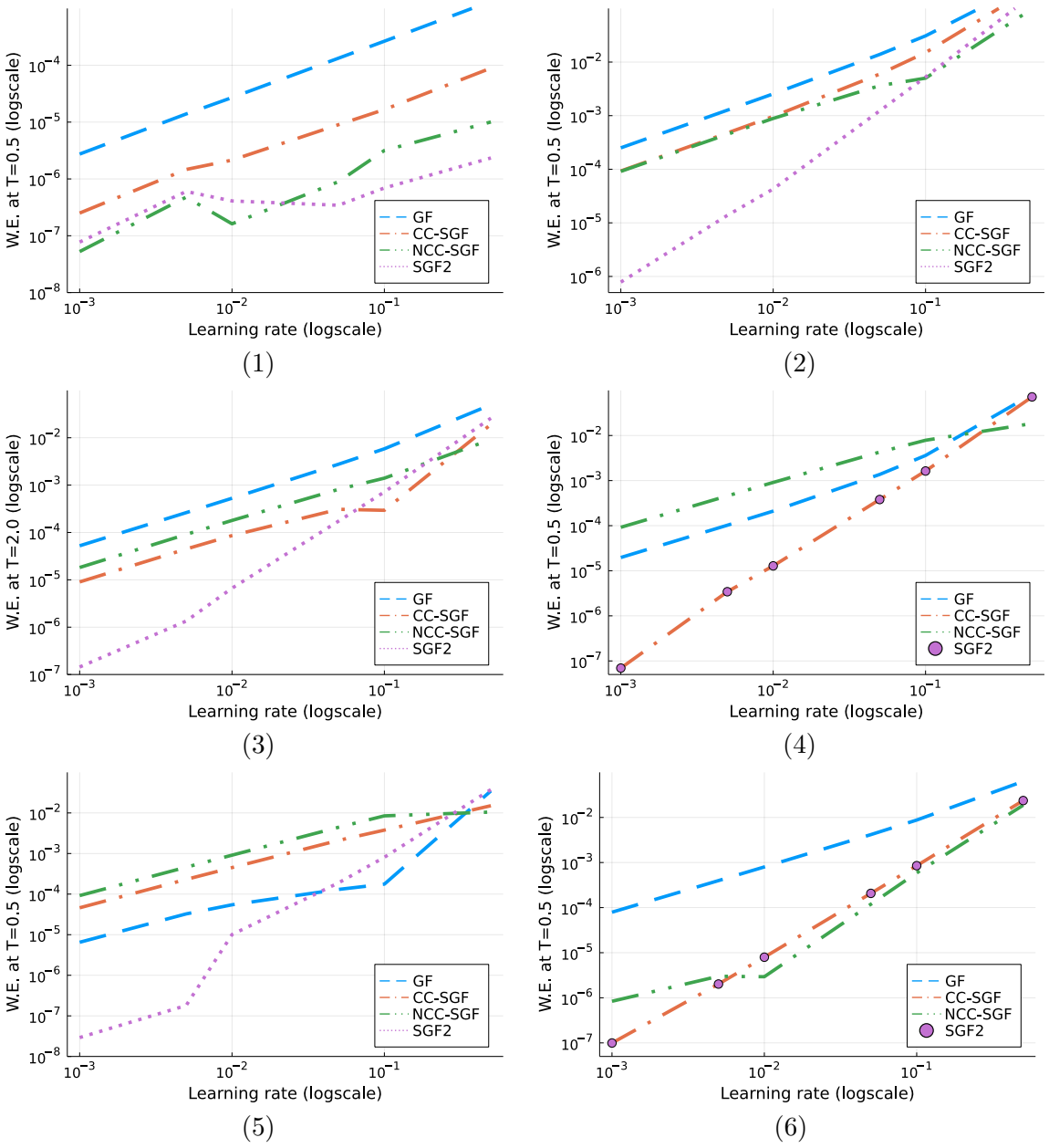

Figure 2: The weak error's dependence on the learning rate for several continuous-time approximations to SGD, in various settings. The plots (1)-(5) correspond to the settings (i)-(v) in Theorem 6. Further, (4) and (6) also correspond to (vi). Finally, (6) depicts a situation where $X^{\mathrm{CC}} \not\approx X^{\mathrm{NCC}}$, but the weak errors are close to each other since the common initial value is close to the minimum.

albeit with somewhat less simplified terms. However, in this case the minimum[6] of Σ may be different from the minimizer θ^* of the objective function $\mathcal{R}$. In contrast, we assumed that Σ is minimized at θ^*. If this is not the case, then it is possible for $\alpha_T^{\Sigma(\theta^*)}$ in Theorem 4 to have negative eigenvalues. Hence, we may have

$$\mathrm{LE}(X^{\mathrm{CC}}) = \mathrm{LE}(X^{\mathrm{NCC}}) + \frac{b}{B}$$

for some b with $b < 0$ (instead of $b > 0$). Thus, Theorem 6 does not easily generalize to this setting.

Disregarding this issue let us consider more general non-quadratic objective functions $\mathcal{R}$ with convergent gradient flow X^0 and Σ which is minimized at θ^*. By assumption $\nabla^3\mathcal{R} \neq 0$, so $\nabla^2 X^0 \neq 0$ and $\beta_T \neq 0$ (see Theorem 4). In general, neither the sign nor the relative size of $\langle \nabla\mathcal{R}(X_T^0), \beta_T \rangle$ in Equation (16) is known. Unfortunately, this means that Theorem 6 cannot be generalized to this setting either. We leave it to future work to study this term in other practically relevant settings.

4. Proof of Theorems 1, 2 and 3

In this section we give proofs of Theorems 1, 2 and 3. However, before doing that we need to establish a few preliminaries.

4.1 Preliminaries

Let I be a set and $X = (X_t^i)_{i\in I, t\geq 0}$ be an I-indexed family of continuous-time stochastic processes. Given $p \in [1,\infty)$ we define

$$\|X\|_{p,t} = \sup_{i\in I}\left(\mathbb{E}\int_0^t |X_s^i|^p\, ds\right)^{1/p}, \quad \|X^*\|_{p,t} = \sup_{i\in I}\left(\mathbb{E}\sup_{s\in[0,t]}|X_s^i|^p\right)^{1/p}.$$

Although usually X will be $\mathbb{R}^d$-valued and then $|\cdot|$ refers to the Euclidean norm, these definitions naturally extend to $\mathbb{R}^{d_1\times\cdots\times d_r}$-valued processes as well. Similarly, given an I-indexed family of discrete-time stochastic processes X we define

$$\|X^*\|_{p,n} = \sup_{i\in I}\left(\mathbb{E}\max_{n'\in\{0,\ldots,n\}}|X_{n'}^i|^p\right)^{1/p}.$$

Given an I-indexed family of random variables $Y = (Y^i)_{i\in I}$ we also let

$$\|Y\|_p := \sup_{i\in I}(\mathbb{E}|Y^i|^p)^{1/p}.$$

Recall the definition of χ in (5), as well as Assumptions (A1) and (A2). We shall prove growth results concerning stochastic gradient descent. Denote the SGD iterations starting

6. Here, we mean minimum in the sense of the Loewner order. In particular, Σ is minimal at θ^* if $\Sigma(\theta) - \Sigma(\theta^*)$ is positive semi-definite, for all $\theta \in \mathbb{R}^d$.

at time k with initial value $x \in \mathbb{R}^d$ and maximal learning rate $h \in (0,1)$ by $\chi^{h,k}(x)$. Given a discrete process Y indexed by $h \in (0,1)$, for example $Y = \chi$, we write

$$\Delta Y_n^{h,k}(x) := Y_{n+1}^{h,k}(x) - Y_n^{h,k}(x), \tag{25}$$

for all $h \in (0,1), k, n \in \mathbb{N}_0$ with $k \leq n$ and initial values $x \in \mathbb{R}^d$. We let $\Delta Y_n^h := \Delta Y_n^{h,0}$. Observe that $\Delta Y_n^{h,n}(x) = Y_{n+1}^{h,n}(x) - x$.

In order to simplify notation, in this section we often omit the initial condition from χ or the solution X of a given SDE and formulate statements for the mapping from the set of initial conditions $\mathbb{R}^d$ to the collection of random variables $(\chi_n)_n$ or $(X_t)_t$.

Lemma 7 *The following estimates hold true:*

(i) For every $T > 0$ and $p \geq 1$ there exists a constant $C > 0$, such that

$$\sup_{h \in (0,1)} \|\chi^h(x)^*\|_{p, \lfloor \frac{T}{h} \rfloor} \leq C(1 + |x|),$$

for $x \in \mathbb{R}^d$.

(ii) There exists a constant $C > 0$, such that

$$\|\Delta \chi_n^{h,n}(x)\|_p \leq hC(1 + |x|),$$

for all $h \in (0,1)$, $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$.

Proof

(i) Let $p \in \mathbb{N}$. For every $h \in (0,1)$ and $n \in \mathbb{N}_0$,

$$\|(\chi^h)^*\|_{p,n} = \left(\mathbb{E} \max_{n' \in \{-1, \dots, n-1\}} |\chi_{n'+1}^h|^p \right)^{1/p}.$$

If we let $\chi_{-1} = 0$, then

$$\begin{aligned} |\chi_{n+1}^h|^p \leq & |\chi_n^h + h u_{nh} H_{\gamma(n)}(\chi_n^h)|^p \\ \leq & |\chi_n^h|^p + \sum_{i=1}^{p} \binom{p}{i} |\chi_n^h|^{p-i} (h u_{nh})^i |H_{\gamma(n)}(\chi_n^h)|^i. \end{aligned}$$

Now, for $i \in \{1, \dots p\}$, $h \in (0,1)$ and $n \in \mathbb{N}_0$,

$$\begin{aligned} \|(|\chi^h|^{p-i} |H_{\gamma(0)}(\chi^h)|^i)^*\|_{1,n} &\leq \|(|\chi^h|^{p-i} \|H\|_{G_1}^i (1 + |\chi^h|)^i)^*\|_{1,n} \\ &\leq \frac{1}{2} c^i \|(|\chi^h|^{p-i} + |\chi^h|^{i+p-i})^*\|_{1,n} \\ &\leq c^i (1 + \|(\chi^h)^*\|_{p,n}^p), \end{aligned}$$

with $c := 2\|H\|_{G_1}$ and using the inequalities $y^p + y^q \leq 2(1+y^q)$ for $0 < p \leq q$ and $y \geq 0$. Therefore,

$$\begin{aligned}
\|(\chi^h)^*\|^p_{p,n+1} \leq & \mathbb{E} \max_{n' \in \{-1,\dots,n\}} |\chi^h_{n'}|^p \\
& + \mathbb{E} \max_{n' \in \{-1,\dots,n\}} \sum_{i=1}^{p} \binom{p}{i} (hu_{n'h})^i |\chi^h_{n'}|^{p-i} |H^h_{\gamma(n')}(\chi^h_{n'})|^i \\
\leq & \|(\chi^h)^*\|^p_{p,n} + \sum_{i=1}^{p} \binom{p}{i} \|((hu_{n'h})^i |\chi^h_{n'}|^{p-i} |H^h_{\gamma(n')}(\chi^h_{n'})|^i)^*\|_{1,n} \\
\leq & \|(\chi^h)^*\|^p_{p,n} + Ch(1 + \|(\chi^h)^*\|^p_{p,n}) \\
= & (1+Ch)\|(\chi^h)^*\|^p_{p,n} + Ch,
\end{aligned}$$

where $C := \sum_{i=1}^{p} \binom{p}{i} c^i$. By induction over n,

$$\|(\chi^h)^*\|^p_{p,n} \leq (1+Ch)^n \|(\chi^h)^*\|^p_{p,0} + Ch \left(\sum_{i=0}^{n-1} (1+Ch)^i \right),$$

for all $h \in (0,1)$ and $n \in \mathbb{N}$. Consequently,

$$\begin{aligned}
\|\chi^h(x)^*\|^p_{p,\lfloor \frac{T}{h} \rfloor} &\leq (1+Ch)^{\lfloor \frac{T}{h} \rfloor} |x|^p + Ch \sum_{i=0}^{\lfloor \frac{T}{h} \rfloor} (1+Ch)^i \\
&\leq (1+Ch)^{\frac{T}{h}} |x|^p + Ch \frac{T}{h} (1+Ch)^{\frac{T}{h}} \\
&= (CT + |x|^p) e^{\log(1+Ch)\frac{T}{h}} \\
&\leq (CT + |x|^p) e^{CT},
\end{aligned}$$

for all $h \in (0,T)$ and $x \in \mathbb{R}^d$, since $\log(1+y) \leq y$ for all $y > -1$. Now, the inclusion follows for $p \in \mathbb{N}$. For arbitrary $p \geq 1$ we have $\|Y^*\|_p \leq \|Y^*\|_{\lceil p \rceil}$ and thus the result is proven.

(ii) We have

$$\|\Delta\chi^{h,n}_n(x)\|_p = \|hu_{nh}H(x)\|_p \leq h\|H\|_{G_1}(1+|x|),$$

for all $x \in \mathbb{R}^d$ and $h \in (0,1)$.

■

We shall now consider moments and growth conditions for solutions of (families of) stochastic differential equations that will act as approximations to SGD. Let $l \in \mathbb{N}_0$. We write $f \in \mathrm{Lip}^l$ if $f \in C^l([0,T] \times \mathbb{R}^d)$ and there exists a $C > 0$ such that

$$|\partial_\alpha f_t(x) - \partial_\alpha f_t(y)| \leq C|x-y|,$$

for all $t \geq 0$ and multi-indices α with size $\#\alpha \leq l$. Also set $\mathrm{Lip} := \mathrm{Lip}^0$. Given an index set I, these conditions extend to I-indexed families of functions $(f_i)_{i \in I}$ in a uniform sense.

Further, we extend the use of the notation G to *families* of functions. More precisely, given a family of functions

$$f : I \times \mathbb{R}^d \to \mathbb{R}, (i, x) \mapsto f_i(x),$$

we write $f \in G(\mathbb{R}^d)$ whenever there exists a constant $C > 0$ and $\kappa \in \mathbb{N}$ such that

$$|f_i(x)| \leq C(1 + |x|^\kappa), \tag{26}$$

for all $x \in \mathbb{R}^d$ and $i \in I$. Again, we define $\|g\|_{G_\kappa}$ as the infimum of all C's in (26).

Notice that the index set may comprise the time interval $[0, T]$. Usually, we have $I = \mathcal{H}$ or or $I = \mathcal{H} \times [0, T]$ or $I = (0, 1)$.

Similarly we extend the use of the notations G^l to families of functions. In particular, for an I-indexed family of functions $f : I \times [0, T] \times \mathbb{R}^d \to \mathbb{R}$ we write $f \in G^\infty([0, T] \times \mathbb{R}^d)$ if each f_i is infinitely continuously differentiable in time and space, and all derivatives have at most polynomial growth, uniformly in $i \in I$. Finally, all the definitions extend naturally to other codomains such as $\mathbb{R}^d$ or $\mathbb{R}^{d\times d}$.

We shall consider stochastic differential equations with (families of) coefficients

$$b : I \times [0, T] \times \mathbb{R}^d \to \mathbb{R}^d, \sigma : I \times [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}.$$

Proposition 8 *Let $l \in \mathbb{N}, p \geq 1$ and $b, \sigma \in G_1(\mathbb{R}^d) \cap \mathrm{Lip}^l$, such that b is $\mathbb{R}^d$-valued and σ is $\mathbb{R}^{d\times d}$-valued. Let X be the unique solution to the family of stochastic differential equations*

$$dX_t^{i,s}(x) = b_t^i(X_t^{i,s}(x))\, dt + \sigma_t^i(X_t^{i,s}(x))\, dW_t, \quad X_s^{i,s}(x) = x.$$

and $g : I \times \mathbb{R}^d \to \mathbb{R} \in G^l(\mathbb{R}^d)$. Define

$$v_t^{i,s}(x) := \mathbb{E}g^i(X_t^{i,s}(x)).$$

Then $v \in G^l(\mathbb{R}^d)$.

Note that the polynomial growth of v and its partial derivatives up to order l is considered uniformly in $i \in I$ and $s, t \in [0, T]$.

Proof Let α be a multi-index. By induction one can show $\mathbb{E}\partial_\alpha g(X) = \partial_\alpha \mathbb{E}g(X)$ using Theorem 27 in the Appendix. By the higher chain rule,

$$|\partial_\alpha v_t^{i,s}| = \mathbb{E}|\partial_\alpha g^i(X_t^{i,s})| \leq \sum_{j=1}^{\#\alpha} \|\nabla^j g^i(X)^*\|_2 \sum_{\mathcal{B}\in\mathcal{S}_j^\alpha} N(\alpha, \mathcal{B}) \prod_{\beta\in\mathcal{B}} \|\partial_\beta X^*\|_{2\#\mathcal{B}},$$

where $\mathcal{S}_i^\alpha$ is the set of all partitions of α into i multi-set multi-indices (each partition being a multi-set as well), $N(\alpha, \mathcal{B}) \in \mathbb{N}$, $\#\mathcal{B}$ is the size of the partition and the product $\prod_{\beta\in\mathcal{B}}$ respects the multiplicities of $\beta \in \mathcal{B}$. From $g \in G^l(\mathbb{R}^d)$ and Theorem 27 we conclude $\partial_\alpha v \in G(\mathbb{R}^d)$. ■

Remark 9 *Assume now we are given an SDE with separable coefficients, specifically*

$$dX_t = u_t B(X_t)\,dt + u_t S(X_t)\,dW_t,$$

where $B, S \in \mathrm{Lip} \cap G^\infty$. Further, suppose Assumption (A1) holds. Given $g \in G^\infty(\mathbb{R}^d)$ we want to show that v defined by

$$v_t^{i,h} := \mathbb{E} g^i(X_T^{h,t})$$

satisfies $v \in G^\infty([0,T] \times \mathbb{R}^d)$.

To this end let U be a map from the image of u to $\mathbb{R}$, such that

$$U = \begin{cases} \dot{u} \circ u^{-1}, & u \text{ strictly monotone}, \\ 0, & u \text{ constant}. \end{cases}$$

Then U is continuous, bounded and

$$du_t = U(u_t)\,dt, t \geq 0.$$

Consider the system

$$dZ_t = b(Z_t)\,dt + \Sigma(Z_t)\,dW_t,$$

with

$$Z_t = \begin{pmatrix} X_t \\ u_t \end{pmatrix}, b\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} yB(x) \\ U(y) \end{pmatrix}, \Sigma\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} yS(x) \\ 0 \end{pmatrix}.$$

Then $b, \Sigma \in G(\mathbb{R}^d)$. If the coefficients of an autonomous SDE

$$dZ_t = b(Z_t)\,dt + \Sigma(Z_t)\,dW_t$$

are in G^∞ and $g \in G^\infty(\mathbb{R}^d)$, then clearly also $L_Z g \in G^\infty([0,T] \times \mathbb{R}^d)$, where L_Z is the infinitesimal generator of Z. By Proposition 8 then $\mathbb{E} L_Z g(Z) \in G^\infty([0,T] \times \mathbb{R}^d)$. If $g \in G^\infty(\mathbb{R}^d)$, then $v_t^{i,h} := \mathbb{E} g^i(X_T^{h,t})$ satisfies the Feynman-Kac equation[7]

$$\partial_t v_t + L_X v_t = 0, \quad v_T = g,$$

where L_X^h is the infinitesimal generator of X^h. In particular,

$$\partial_t \mathbb{E} g(X_T^t) = \partial_t \mathbb{E} g(Z_T^t) = \partial_t \mathbb{E} g(Z_{T-t}^0) = L_Z(\mathbb{E} g(Z_{T-t}^0)) \in G([0,T] \times \mathbb{R}^d),$$

with the understanding that $g(x,y) := g(x)$. Inductively,

$$\partial_\alpha \partial_t^k \mathbb{E} g(X_T^t) = \partial_\alpha L_Z^k \mathbb{E}(g(Z_{T-t}^0)) \in G([0,T] \times \mathbb{R}^d).$$

All in all we have $v \in G^\infty([0,T] \times \mathbb{R}^d)$, that is v is smooth in time and space, and all its derivatives have polynomial growth (uniformly in time).

7. See, for example, the book by Graham and Talay (2013, Theorem 7.14 and Remark 7.6).

Next we shall consider *families* of stochastic differential equations

$$dX_t^h = b_t^h(X_t^h)\,dt + \sqrt{h}\sigma_t^h(X_t^h)\,dW_t,$$

indexed by a discretization parameter $h \in (0,1)$. Given the family of solutions X of an h-indexed family of stochastic differential equations we define the family of discrete processes

$$\tilde{X}_n^h(x) := X_{nh}^h(x), \tag{27}$$

with $h \in (0,1), x \in \mathbb{R}^d$ and $n \in \{0, \dots, \lfloor T/h \rfloor\}$. Then,

$$\Delta\tilde{X}_n^{h,n}(x) = X_{nh}^h(x) - x.$$

Lemma 10 *Let*

$$b : (0,1) \times [0,T] \times \mathbb{R}^d \to \mathbb{R}^d, \sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d} \in G_1(\mathbb{R}^d) \cap \mathrm{Lip}$$

and X be the unique solution to family of stochastic differential equations

$$dX_t^h = b_t^h(X_t^h)\,dt + \sqrt{h}\sigma_t(X_t^h)\,dW_t, \quad X_0^h(x) = x \in \mathbb{R}^d, h \in (0,1).$$

Then for all $p \geq 2$ there exists a $C \in G(\mathbb{R}^d)$, such that

$$\|\Delta\tilde{X}_n^{h,n}\|_p \leq hC,$$

for all $h \in (0,1)$ and $n \in \{0, \dots, \lfloor T/h \rfloor\}$.

Proof We have

$$\|\Delta\tilde{X}_n^{h,n}\|_p \leq \| \int_{nh}^{(n+1)h} b_s^h(X_s)ds\|_p + \sqrt{h}\| \int_{nh}^{(n+1)h} \sigma(X_s^h)\,dW_s\|_p.$$

On the one hand

$$\begin{aligned}
\| \int_{nh}^{(n+1)h} b_t^h(X_t^h)dt\|_p \leq & h^{1-\frac{1}{p}} \left(\int_{nh}^{(n+1)h} \mathbb{E}|b_t^h(X_t)|^p\,dt \right)^{1/p} \\
\leq & h \left(\mathbb{E} \sup_{t,h} |b_t^h(X_t)|^p \right)^{1/p} \\
= & h\|b(X)^*\|_p,
\end{aligned}$$

and $x \mapsto \|b(X(x))^*\|_p \in G(\mathbb{R}^d)$ by Theorem 26 and since $b \in G_1(\mathbb{R}^d)$. On the other hand,

$$\begin{aligned}
\sqrt{h}\| \int_{nh}^{(n+1)h} \sigma_t(X_t^h)\,dW_t\|_p \leq & \sqrt{\frac{p(p-1)}{2}} h^{1-\frac{1}{p}} \|\sigma(X^h)\|_p \\
\leq & c_1 h\|\sigma(X)^*\|_p,
\end{aligned}$$

where we have used Itô's isometry and Jensen's inequality. ∎

Lemma 11 *Let $b, \sigma \in G_1([0,\infty)\times\mathbb{R}^d)\cap G^\infty([0,\infty)\times\mathbb{R}^d)$, such that b is $\mathbb{R}^d$-valued and σ is $\mathbb{R}^{d\times d}$-valued. Let $h\in(0,1), s\geq 0$ and consider the stochastic differential equation*

$$dX_t^h = b_t(X_t^h)\,dt + \sqrt{h}\sigma_t(X_t^h)\,dW_t, \quad X_s = x,$$

with $t\in[s,s+h]$. Then there exists a function $C\in G(\mathbb{R}^d)$, such that

$$\mathbb{E}[\Delta X_s^h] = hb_s + h^2 C, \qquad \mathbb{E}[(\Delta X_s^h)^{\otimes 2}] = h^2 C, \tag{28}$$

for all $h\in(0,1)$, where $\Delta X_s^h := X_{s+h}^h - X_s^h$.

Proof For any multi-index α define

$$m_\alpha(z) := (z-x)^\alpha = \prod_{j=1}^d (z_j - x_j)^{\alpha(j)}.$$

Then for any other multi-index β,

$$\partial_\beta m_\alpha(z) = \prod_{j=1}^d \prod_{k=1}^{\beta(j)} (\alpha(j)-k+1)(z-x)^{\alpha-\beta}, z\in\mathbb{R}^d,$$

where it is understood that $y^{\alpha-\beta}=0$ if $\alpha(j)<\beta(j)$ for any $j\in\{1,\dots,d\}$. Further, $(\Delta X_s^h)^\alpha = m_\alpha(X_{s+h}^h)$. Write

$$\mathcal{A}_X g = \partial_t g + b^\dagger \nabla g + \frac{h}{2}\operatorname{tr}[\sigma^\dagger\sigma\nabla^2 g], \quad g\in G^\infty,$$

and $\mathcal{A}_X^2 = \mathcal{A}_X\circ\mathcal{A}_X$. Observe that $\mathcal{A}_X g$ already depends on time even if g does not. An Itô-Taylor expansion implies (see Theorem 30)

$$\mathbb{E}(\Delta X_s^h(x))^\alpha = h\mathcal{A}_X m_\alpha(s,x) + \int_s^{s+h}\int_s^t \mathbb{E}\mathcal{A}_X^2 m_\alpha(u, X_u^h(x))\,du\,dt, \quad x\in\mathbb{R}^d, h\in(0,1).$$

We have

$$\mathcal{A}_X(m_j)(s,x) = b_s(x)_j.$$

Moreover, by Lemma 31, $\mathcal{A}_X^2 m_j \in G([0,T]\times\mathbb{R}^d)$, so Theorem 26 implies

$$\|(\mathcal{A}_X^2 m_j(s, X_s^h(x)))\|_1 \leq C(1+\||X_s^h(x)|^\kappa\|_1) \leq C(1+|x|^\kappa), \quad x\in\mathbb{R}^d, h\in(0,1),$$

for some constant $C>0$. Hence,

$$\mathbb{E}[\Delta X_s^h] = hb_s + h^2 C,$$

for some $C\in G(\mathbb{R}^d)$. Now, let us consider a multi-index $\alpha=\{j_1,j_2\}$. Then,

$$\mathcal{A}_X(m_\alpha)(s,x) = \frac{h}{2}(\sigma^\dagger\sigma)_s(x_{j_1}x_{j_2}),$$

with $\sigma\in G$. Again using Lemma 31 we can estimate the remainder term to arrive at

$$\mathbb{E}[(\Delta X_s^h)^{\otimes 2}] = h^2 C,$$

for some $C\in G(\mathbb{R}^d)$, for all $h\in(0,1)$. ■

4.2 Proof of the Gradient Flow Approximation

We shall give a proof of Theorem 1. Fix $g \in G^\infty(\mathbb{R}^d)$ and define once more $v_t(x) := g(X_T^{0,t}(x))$, where X^0 is the solution to the gradient flow equation (6),

$$dX_t^0 = u_t \bar{H}(X_t^0)\, dt.$$

We then have $v \in G^\infty([0,T] \times \mathbb{R}^d)$ by Proposition 8 and Remark 9 and since we have $\bar{H} \in G^\infty(\mathbb{R}^d)$ by Assumption (A3). Further, v satisfies the *Feynman-Kac equation*

$$\partial_t v_t(x) + \nabla v_t(x)^\dagger u_t \bar{H}(x) = 0, \quad v_T(x) = g(x). \tag{29}$$

From now on let χ^h and X^0 denote the solutions of (5) and (6), respectively, with the same fixed initial condition $\chi_0 \in \mathbb{R}^d$.

Recall the definition of φ in (10) and the statement of Theorem 1. We define

$$\varphi_t^{\mathrm{GF}}(x) = \varphi_t(x) + \frac{1}{2} u_t^2 \operatorname{tr}[\nabla^2 v_t(x) \Sigma(x)],$$

for all $x \in \mathbb{R}^d$ and $t \in [0,T]$.

Lemma 12 *Let $\xi : \mathcal{H} \to \mathbb{R}$ be the function such that for all $h \in \mathcal{H}$*

$$\mathbb{E} g(\chi_{T/h}^h) - g(X_T^0) = h^2 \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E} \varphi_{kh}^{\mathrm{GF}}(\chi_k^h) + h^2 \xi(h).$$

Then ξ is bounded.

Proof By Taylor's theorem,

$$\begin{aligned} v_{t+h}(x+\delta) - v_t(x) =& h\partial_t v_t(x) + \nabla v_t(x)^\dagger \delta + \frac{h^2}{2} \partial_t^2 v_t(x) \\ &+ h \partial_t \nabla v_t(x)^\dagger \delta + \frac{1}{2} \operatorname{tr}[\nabla^2 v_t(x) \delta^{\otimes 2}] \\ &+ r^h(\delta), \end{aligned}$$

where

$$r^h(\delta) := \sum_{k=0}^{3} \sum_{\#\beta = 3-k} \frac{1}{\beta! k!} \partial_t^k \partial_\beta v_{t+\theta h}(x + \theta\delta) h^k \delta^\beta$$

for some $\theta \in (0,1)$, all $h \in (0,1)$ and $\delta \in \mathbb{R}^d$. By choosing $t = kh$, $x = \chi_k^h$, $\delta = \Delta\chi_k^h$ and applying expectation we get

$$\mathbb{E} v_{(k+1)h}(\chi_{k+1}^h) - \mathbb{E} v_{kh}(\chi_k^h) = h A_1^h + h^2 (A_2^h + A_3^h + A_4^h) + \mathbb{E} r^h(\Delta\chi_k^h),$$

where

$$\begin{aligned} A_1^h &:= \mathbb{E}[\partial_t v_{kh}(\chi_k^h) + h^{-1} \nabla v_{kh}(\chi_k^h)^\dagger \Delta\chi_k^h], \\ A_2^h &:= \frac{1}{2} u_{kh}^2 \mathbb{E} \operatorname{tr}[\nabla^2 v_{kh}(\chi_k^h)((\bar{H}(\chi_k^h) + (H_{\gamma(0)} - \bar{H})(\chi_k^h))^{\otimes 2})], \\ &= \frac{1}{2} u_{kh}^2 \mathbb{E} \operatorname{tr}[\nabla^2 v_{kh}(\chi_k^h)(\bar{H}^{\otimes 2} + \Sigma)(\chi_k^h)] \\ A_3^h &:= u_{kh} \mathbb{E}[\partial_t \nabla v_{kh}(\chi_k^h)^\dagger \bar{H}(\chi_k^h)], \\ A_4^h &:= \frac{1}{2} \mathbb{E}[\partial_t^2 v_{kh}(\chi_k^h)]. \end{aligned}$$

Using the Feynman-Kac equation (29) we can simplify

$$A_1^h = \mathbb{E}[\mathbb{E}(\partial_t v_{kh}(\chi_k^h) + \nabla v_{kh}(\chi_k^h)^\dagger u_{kh} \bar{H}(\chi_k^h)|\chi_k^h)] = 0.$$

We want to show that the remainder satisfies $\mathbb{E}r^h(\Delta\chi_n^h) = \mathcal{O}(h^3)$. For $k \in \{0,\dots,3\}$ and $\#\beta = 3-k$,

$$\mathbb{E}[h^k(\Delta\chi_n^h)^\beta] = h^k h^{3-k}(u_{nh})^{3-k}\mathbb{E}\bar{H}(\chi_n^h)^\beta = \mathcal{O}(h^3),$$

since $u \le 1$ and

$$\begin{aligned}\mathbb{E}[|\bar{H}(\chi_n^h)^\beta|]^{1/\#\beta} \le& \sup_{h\in(0,1)} \|\bar{H}(\chi^h)^*\|_{\#\beta,\lfloor\frac{T}{h}\rfloor} \\ \le& \|\bar{H}\|_{G_1}\left(1 + \sup_{h\in(0,1)} \|(\chi^h)^*\|_{\#\beta,\lfloor\frac{T}{h}\rfloor}\right) \\ \le& c(1+|\chi_0|),\end{aligned}$$

by Lemma 7. Since $\partial_t^k\partial_\alpha^{2-k} v \in G([0,T]\times\mathbb{R}^d)$ for all $k \in \{0,1,2\}$, we have $\mathbb{E}r^h(\Delta\chi_n^h) = \mathcal{O}(h^3)$. Therefore,

$$\begin{aligned}\mathbb{E}g(\chi_{T/h}^h) - g(X_T^0) =& \mathbb{E}v_T(\chi_{T/h}^h) - \mathbb{E}v_0(\chi_0) \\ =& \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E}v_{(k+1)h}(\chi_{k+1}^h) - \mathbb{E}v_{kh}(\chi_k^h) \\ =& h^2 \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E}\varphi_{kh}^{\mathrm{GF}}(\chi_k^h) + \mathcal{O}(h^2),\end{aligned}$$

for all $h \in \mathcal{H}$. ∎

The bound on the function ξ in Lemma 12 only depends on the growth of g and its derivatives, as well as $\bar{H}$, Σ and T. We use this fact in the next step, where we apply Lemma 12 to the family of functions $(\varphi_{nh}^{\mathrm{GF}})_{h\in\mathcal{H},n\le T/h}$.

For all $h \in \mathcal{H}$ and $n \in \{0,\dots,T/h\}$, let $\xi_n(h) \in \mathbb{R}$ be, such that

$$\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) - \varphi_{nh}^{\mathrm{GF}}(X_{nh}^h) = h^2 \sum_{k=0}^{n-1} \mathbb{E}\psi_{nh,kh}(\chi_k^h) + h^2\xi_n(h) \tag{30}$$

with

$$\begin{aligned}\psi_{s,t}(x) :=& \frac{1}{2}u_t^2 \operatorname{tr}[\nabla^2 z_{s,t}(x)(\bar{H}^{\otimes 2} + \Sigma)(x)] + u_t\partial_t\nabla z_{s,t}(x)\bar{H}(x) \\ &+ \frac{1}{2}\partial_t^2 z_{s,t}(x), \\ z_{s,t} :=& \varphi_s^{\mathrm{GF}}(X_s^{0,t}).\end{aligned}$$

Now choose a constant $B \in [0,\infty)$ such that for all n and h we have

$$|\xi_n(h)| \le B. \tag{31}$$

Using this estimate we can bound the differences of the form $\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) - \varphi_{nh}^{\mathrm{GF}}(X_{nh})$.

Lemma 13 *There exists a constant $C > 0$ such that*

$$\sum_{n=0}^{\frac{T}{h}-1} |\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) - \varphi_{nh}^{\mathrm{GF}}(X_{nh}^0)| \leq C$$

for all $h \in \mathcal{H}$.

Proof By (30) and (31)

$$\begin{aligned}\sum_{n=0}^{\frac{T}{h}-1} |\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) - \varphi_{nh}^{\mathrm{GF}}(X_{nh}^0)| \leq& h^2 \sum_{n=0}^{\frac{T}{h}-1}\sum_{k=0}^{n-1} \mathbb{E}|\psi_{nh,kh}(\chi_k^h)| + Bh \\ \leq& C\left(1 + \max_{n,k} \mathbb{E}|\psi_{nh,kh}(\chi_k^h)|\right),\end{aligned}$$

for some $C > 0$ and all $h \in (0,1)$.

Because $\partial_t^k \partial_\alpha^{2-k} v \in G([0,T]\times\mathbb{R}^d)$ for all $k \in \{0,1,2\}$, $g \in G(\mathbb{R}^d)$, u is bounded and $\bar{H}, \Sigma \in G(\mathbb{R}^d)$, we have $\varphi^{\mathrm{GF}} \in G([0,T]\times\mathbb{R}^d)$. With Lemma 7,

$$\begin{aligned}\max_{n,k} \mathbb{E}|\psi_{nh,kh}(\chi_n^h)| \leq& \|\varphi^{\mathrm{GF}}\|_{G_\kappa}\left(1 + \sup_{h\in(0,1)} \|(\chi^h)^*\|_1^\kappa\right) \\ \leq& C(1 + |\chi_0|^\kappa),\end{aligned}$$

for some $C > 0, \kappa \in \mathbb{N}$ and all $h \in (0,1)$. ■

Proof of Theorem 1 Let $g \in G^\infty(\mathbb{R}^d)$ and $h \in \mathcal{H}$. Then Lemma 12 implies

$$\mathbb{E}g(\chi_{T/h}^h) - g(X_T^0) = h\sum_{n=0}^{\frac{T}{h}-1} h\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) + \mathcal{O}(h^2),$$

We can then write the linear error term as follows.

$$\begin{aligned}\sum_{n=0}^{\frac{T}{h}-1} h\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) =& \int_0^T \varphi_t^{\mathrm{GF}}(X_t^0)\,dt + h\sum_{n=0}^{\frac{T}{h}-1} \mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) - \varphi_{nh}^{\mathrm{GF}}(X_{nh}^0) \\ &+ \sum_{n=0}^{\frac{T}{h}-1} h\varphi_{nh}^{\mathrm{GF}}(X_{nh}^0) - \int_0^T \varphi_t^{\mathrm{GF}}(X_t^0)\,dt,\end{aligned}$$

Using Lemma 13, we then have

$$h\sum_{n=0}^{\frac{T}{h}-1} |\mathbb{E}\varphi_{nh}^{\mathrm{GF}}(\chi_n^h) - \varphi_{nh}^{\mathrm{GF}}(X_{nh}^0)| \leq hC.$$

Further, approximating the integral $\int \varphi^{\mathrm{GF}}$ by a left Riemann sum yields

$$\left| \sum_{n=0}^{\frac{T}{h}-1} h\varphi_{nh}^{\mathrm{GF}}(X_{nh}^0) - \int_0^T \varphi_t^{\mathrm{GF}}(X_t^0)\, dt \right| \leq hC'.$$

Hence,

$$\mathbb{E}g(\chi_{T/h}^h) - g(X_T^0) = h\int_0^T \varphi_t^{\mathrm{GF}}(X_t^0)\, dt + \mathcal{O}(h^2),$$

for all $h \in \mathcal{H}$. ■

4.3 Proof of the Stochastic Gradient Flow Approximations

The first part of the proofs of Theorems 2 and 3 are somewhat analogous to the ODE case. We focus on proving Theorem 2 while omitting the proof of 3 since it is completely analogous. One notable difference to the GF case comes from the newly acquired dependence of the solution $X := X^{\mathrm{NCC}}$ of (12) on $h \in \mathcal{H}$. This carries over to v and by extension to the function

$$\varphi_t^h(x) := \frac{1}{2}u_t^2 \operatorname{tr}[\nabla^2 v_t^h(x)\bar{H}^{\otimes 2}(x)] + u_t\partial_t\nabla v_t^h(x)\bar{H}(x) + \frac{1}{2}\partial_t^2 v_t^h(x).$$

Note the absence of the Σ term compared to the ODE case. By using arguments as in Section 4.2, we arrive at an approximation of the form

$$\mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^h) = h\int_0^T \mathbb{E}\varphi_t^h(X_t^h)\, dt + \mathcal{O}(h^2). \tag{32}$$

We then need to improve the estimate to

$$\mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^h) = h\int_0^T \varphi_t^0(X_t^0)\, dt + \mathcal{O}(h^2).$$

This requires an additional estimation of the difference $\varphi_t^h(X_t^h) - \varphi_t^0(X_t^0)$. Let us be more specific now. Let $g \in G^\infty(\mathbb{R}^d)$ and define, for all $h \in [0,1)$, $t \in [0,T]$ and $x \in \mathbb{R}^d$,

$$v_t^h(x) := \mathbb{E}g(X_T^{h,t}(x)),$$

where $X^{h,t}(x)$ denotes the solution of (12) on $[t,T]$ with initial condition $X_t^{h,t}(x) = x$. Then $v \in G^\infty([0,T]\times\mathbb{R}^d)$, as defined in (26) with $I = \mathcal{H}$, and it satisfies the Feynman-Kac equation

$$\partial_t v_t(x) + \nabla y_t^\dagger(x)u_t\bar{H}(x) + \frac{1}{2}hu_t^2 \operatorname{tr}[\nabla^2 v_t(x)\Sigma(x)] = 0, \quad v_T(x) = g(x). \tag{33}$$

Given a family $(f_t^h)_{h\in(0,1),t\geq 0}$ of continuous-time stochastic processes (or merely functions) we define for every $h \in (0,1)$ the discrete-time process

$$\tilde{f}_n^h := f_{nh}^h, n \in \mathbb{N}.$$

From now on let χ^h and X^h denote the solutions of (5) and (12), respectively, with the same fixed initial condition $\chi_0 \in \mathbb{R}^d$ and $h \in \mathcal{H}$. Then we have the following.

Lemma 14 *We have*

$$\mathbb{E}g(\chi^h_{T/h}) - \mathbb{E}g(X^h_T) = h^2 \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E}\Phi^h_k(\chi^h_k) + \mathcal{O}(h^2),$$

for all $h \in \mathcal{H}$*, where* $\Phi^h := \tilde{\varphi}^h$*.*

Proof Follow the proof of Lemma 12. A Taylor expansion of v yields

$$\mathbb{E}\tilde{v}^h_{k+1}(\chi^h_{k+1}) - \mathbb{E}\tilde{v}^h_k(\chi^h_k) = hA^h_1 + h^2(A^h_2 + A^h_3 + A^h_4) + \mathbb{E}r^h(\Delta\chi^h_k),$$

as before, except with

$$\begin{aligned} A^h_1 :=& \mathbb{E}[\partial_t \tilde{v}^h_k(\chi^h_k) + h^{-1}\nabla\tilde{v}^h_k(\chi^h_k)^\dagger \Delta\chi^h_k + \frac{1}{2}hu^2_{kh}\operatorname{tr}[\nabla^2\tilde{v}^h_k(\chi^h_k)\Sigma(\chi^h_k)]] \\ =& 0 \end{aligned}$$

by (33) and to compensate for the additional term

$$A^h_2 := \frac{1}{2}u^2_{kh}\mathbb{E}\operatorname{tr}[\nabla^2\tilde{v}^h_k(\chi^h_k)\bar{H}^{\otimes 2}(\chi^h_k)].$$

■

Again, we could have stated Lemma 14 with g depending on h and t, so the following holds.

Lemma 15 *With the conditions as in Lemma 14, there exists a constant* $C > 0$ *with*

$$\sum_{n=0}^{\frac{T}{h}-1} |\mathbb{E}\Phi^h_n(\chi^h_n) - \mathbb{E}\Phi^h_n(\tilde{X}^h_n)| \le C$$

for all $\mathcal{H} \ni h \downarrow 0$*.*

Our initial approximation follows just as in the ODE case, so we shall omit the proof of the following lemma.

Lemma 16 *For all* $g \in G^\infty(\mathbb{R}^d)$ *and* $h \in \mathcal{H}$*,*

$$\mathbb{E}g(\chi^h_{T/h}) - \mathbb{E}g(X^h_T) = h\int_0^T \mathbb{E}\varphi^h_t(X^h_t)\,dt + \mathcal{O}(h^2), \tag{34}$$

where

$$\varphi^h_t(x) = \frac{1}{2}u^2_t \operatorname{tr}[\nabla^2 v^h_t(x)\bar{H}^{\otimes 2}(x)] + u_t\partial_t\nabla v^h_t(x)\bar{H}(x) + \frac{1}{2}\partial^2_t v^h_t(x).$$

Next we shall improve (34) in order to arrive at the equality in Theorem 2. An additional step compared to the ODE approximation is then deriving an estimate of $|\mathbb{E}\varphi^h_t(X^h_t) - \varphi^h_0(X^0_t)|$ to get rid of the dependence of the integral $\int_0^T |\mathbb{E}\varphi^h_t(X^h_t)|\,dt$ on $h \in (0,1)$. First, consider estimating the difference $v^h - v^0$ and its derivatives up to order 2.

Lemma 17 *Let $v_t^h(x) = \mathbb{E}g(X_T^{h,t}(x))$. Define the $\mathcal{H}$-indexed family*

$$\delta_t^h(x) := \frac{v_t^h(x) - v_t^0(x)}{h}.$$

Then $\delta^h \in G^2([0,T]\times\mathbb{R}^d)$, uniformly in h.

Proof For every $s \in [0,T]$ and $h \in \mathcal{H}$, such that $\frac{s}{h} \in \mathbb{N}_0$ we have

$$|v_s^h - v_s^0| \leq \sum_{n=0}^{\frac{T-s}{h}-1} |\mathbb{E}v^0_{s+(n+1)h}(X^{h,s}_{s+(n+1)h}) - \mathbb{E}v^0_{s+nh}(X^{h,s}_{s+nh})|,$$

where this is meant as an inequality of functions on $\mathbb{R}^d$, the set of possible initial values. To shorten notation, throughout this proof we omit the initial value in $X^{h,s}(x)$.

Set $A_t^h := v^0_{t+h}(X^{h,s}_{t+h}) - v_t^0(X_t^{h,s})$. Since $v^0 \in G^\infty([0,T]\times\mathbb{R}^d)$, applying Taylor's theorem to it implies

$$\begin{aligned} A_t^h =& \partial_t v_t^0(X_t^{h,s})h + \nabla v_t^0(X_t^{h,s})^\dagger \Delta X_t^{h,s} + \frac{1}{2}\operatorname{tr}[\nabla^2 v_t^0(X_t^{h,s})(\Delta X_t^{h,s})^{\otimes 2}] \\ &+ h^2 r_t^h(\Delta X_t^{h,s}) \end{aligned}$$

with some remainder term $r : \mathcal{H}\times[0,T]\times\mathbb{R}^d \to \mathbb{R} \in G([0,T]\times\mathbb{R}^d)$ and $\Delta X_t^{h,s} := X^{h,s}_{t+h} - X_t^{h,s}$. By the Feynman-Kac formula (33),

$$\begin{aligned} \mathbb{E}A_t^h =& \mathbb{E}[\nabla v_t^0(X_t^{h,s})(\Delta X_t^{h,s} - h u_t \bar{H}(X_t^{h,s}))] \\ &+ \frac{1}{2}\mathbb{E}\operatorname{tr}[\nabla^2 v_t^0(X_t^{h,s})((\Delta X_t^{h,s})^{\otimes 2} - h^2 u_t^2 \Sigma(X_t^{h,s}))] + h^2 \mathbb{E} r_t^h(\Delta X_t^{h,s}). \end{aligned}$$

With an Itô-Taylor expansion (see Lemma 11) we see that there exists a $C \in G(\mathbb{R}^d)$ with

$$\begin{aligned} \|\Delta X_t^{h,s} - h u_t \bar{H}(X_t^{h,s})\|_2 \leq& Ch^2, \\ \|(\Delta X_t^{h,s})^{\otimes 2} - h^2 u_t^2 \Sigma(X_t^{h,s})\|_2 \leq& Ch^2, \end{aligned}$$

for all $h \in (0,1)$ and $s,t \in [0,T]$ with $s \leq t$. Since ∇v^0 and $\nabla^2 v^0$ have polynomial growth, uniformly in space and time, there exists a $C \in G(\mathbb{R}^d)$ with

$$\begin{aligned} |\mathbb{E}A_t^h| \leq& \|\nabla v_t^0(X_t^{h,s})\|_2 \|\Delta X_t^{h,s} - h u_t \bar{H}(X_t^{h,s})\|_2 \\ &+ \frac{1}{2}\|\nabla^2 v_t^0(X_t^{h,s})\|_2 \|(\Delta X_t^{h,s})^{\otimes 2} - h^2 u_t^2 \Sigma(X_t^{h,s})\|_2 + h^2 |\mathbb{E} r_t^h(\Delta X_t^{h,s})| \\ \leq& Ch^2, \quad h \in \mathcal{H}, \end{aligned}$$

by Theorem 26 and using the Cauchy-Schwarz inequality. We conclude

$$|v_s^h - v_s^0| \leq \frac{T}{h} Ch^2 \leq TCh,$$

for some $C \in G(\mathbb{R}^d)$, all $h \in \mathcal{H}$ and $s \in [0,T]$ such that $\frac{s}{h} \in \mathbb{N}_0$. For general $t \in [0,T]$ with $nh \leq t < (n+1)h$ a Taylor approximation yields

$$|v_t^h - v_{nh}^h| \leq (t - nh)|\partial_t v_t^h| + h^2 r$$

for some remainder $r \in G([0,T] \times \mathbb{R}^d)$. Since $\partial_t v \in G([0,T] \times \mathbb{R}^d)$ and $(t - nh) \leq h$ we conclude the existence of a $C \in G(\mathbb{R}^d)$ with

$$|v_t^h - v_{nh}^h| \leq Ch,$$

for all $h \in \mathcal{H}$. A similar argument applies to the difference $v_t^0 - v_{nh}^0$. Hence,

$$|v_t^h - v_t^0| \leq |v_t^h - v_{nh}^h| + |v_{nh}^h - v_{nh}^0| + |v_{nh}^0 - v_t^0| \leq Ch,$$

for some $C \in G(\mathbb{R}^d)$, all $h \in \mathcal{H}$ and $t \in [0,T]$. This shows that $\delta^h \in G([0,T] \times \mathbb{R}^d)$, uniformly in h.

Now, we want to show that the partial derivatives of δ up to order 2 have the same property. Fix $j \in \{1, \dots, d\}$ and define

$$w_t^h(x,y) = \mathbb{E}[\nabla g(X_T^{h,t}(x))^\dagger \partial_j X_T^{h,t}(x,y)].$$

Note that $w^h(x,1) = \partial_j v^h(x)$. Furthermore, by differentiating the SDE (12) governing X with respect to its initial condition (see 27), we see that the partial derivative $Y_r := \partial_j X_r^{h,t}(x,y)$ satisfies

$$dY_r = u_r \nabla \bar{H}(X_r^{h,t}(x)) Y_r \, dr + u_r \sqrt{h} \nabla \sqrt{\Sigma(X_r^{h,t}(x))} Y_r \, dW_r,$$

with initial condition $Y_t = y$, where

$$(\nabla \sqrt{\Sigma(x)} y)_{i,j} = \sum_{k=1}^d \partial_i \sqrt{\Sigma(x)_{j,k}} y_k,$$

for all $x, y \in \mathbb{R}^d$ and $i, j \in \{1, \dots, d\}$. The Feynman-Kac equation applies to the system $(X_r^{h,t}, \partial_j X_r^{h,t})$ giving us

$$\begin{aligned} 0 =& \partial_t w_t^h(x,y) + u_t \nabla_x w_t^h(x,y) \bar{H}(x) + \nabla_y w_t^h(x,y) y \partial_j \bar{H}(x) \\ &+ \frac{1}{2} h u_t^2 \operatorname{tr}[\nabla_{x,y}^2 w_t^h(x,y) S(x,y)], \end{aligned}$$

with S given by the block matrix

$$S(x,y) := \begin{pmatrix} \Sigma(x) & \sqrt{\Sigma(x)}(\nabla \sqrt{\Sigma(x)} y)^\dagger \\ \nabla \sqrt{\Sigma(x)} y \sqrt{\Sigma(x)}^\dagger & (\nabla \sqrt{\Sigma(x)} y)(\nabla \sqrt{\Sigma(x)} y)^\dagger \end{pmatrix}.$$

Similarly to the above argument, using Taylor's theorem we can show

$$\begin{aligned} x \mapsto \frac{1}{h} (&\mathbb{E} w_{t+(n+1)h}^0(X_{t+(n+1)h}^h(x), \partial_j X_{t+(n+1)h}^h(x,1)) && (35) \\ &- \mathbb{E} w_{t+nh}^0(X_{t+nh}^h(x), \partial_j X_{t+nh}^h(x,1))) \in G(\mathbb{R}^d) && (36) \end{aligned}$$

and conclude, using a telescoping sum,

$$\frac{1}{h}(\partial_j v_t^h - \partial_j v_t^0) \in G(\mathbb{R}^d).$$

By differentiating the process X once more, an analogous argument works for any second space-derivative to prove

$$\frac{1}{h}|\partial_{i,j}v_t^h - \partial_{i,j}v_t^0| \in G(\mathbb{R}^d),$$

with $i,j \in \{1,\dots,d\}$. Then use the Feynman-Kac equation for v to conclude

$$\frac{1}{h}|\partial_t v_t^h - \partial_t v_t^0| \in G(\mathbb{R}^d).$$

We can then do essentially the same for $\partial_j \partial_t y$ with $j \in \{1,\dots,d\}$ and $\partial_t^2 y$. ■

Consider the linear operator

$$\mathcal{F} : G^2([0,T]\times\mathbb{R}^d) \to G([0,T]\times\mathbb{R}^d)$$

given by

$$\mathcal{F}_t f(x) := \frac{1}{2}u_t^2 \operatorname{tr}(\nabla^2 f_t(x)\bar{H}^{\otimes 2}(x)) + u_t \partial_t \nabla f_t(x)\bar{H}(x) + \frac{1}{2}\partial_t^2 f_t(x).$$

Implicitly, we have already seen it in action. In particular, $\varphi_t^h(x) = \mathcal{F}_t v^h(x)$ for all $t \in [0,T]$ and $x \in \mathbb{R}^d$. In the next lemma we consider spaces of the form

$$G_\kappa^l([0,T]\times\mathbb{R}^d) = \{f \in C^l([0,T]\times\mathbb{R}^d) : \|\partial_t^k \partial_\alpha f\|_{G_\kappa} < \infty, k \le l, |\alpha| \le l-k\}.$$

This is a Banach space when equipped with the norm

$$\|f\|_{G_\kappa^l} := \sum_{k=0}^{l} \sum_{|\alpha|\le l-k} \|\partial_t^k \partial_\alpha f\|_{G_\kappa}.$$

This works regardless of whether we consider functions $f : [0,T]\times\mathbb{R}^d \to \mathbb{R}$ or families of functions, such as $f : \mathcal{H}\times[0,T]\times\mathbb{R}^d \to \mathbb{R}$ with polynomial growth uniformly in $\mathcal{H}$ and $[0,T]$. Of course, by construction

$$G^l([0,T]\times\mathbb{R}^d) = \bigcup_{\kappa\in\mathbb{N}_0} G_\kappa^l([0,T]\times\mathbb{R}^d).$$

Lemma 18 *Let $\kappa \in \mathbb{N}_0$. The function*

$$\mathcal{F} : G_\kappa^2([0,T]\times\mathbb{R}^d) \to G_{\kappa+2}([0,T]\times\mathbb{R}^d)$$

with

$$\mathcal{F}_t f(x) = \frac{1}{2}u_t^2 \operatorname{tr}[\nabla^2 f_t(x)\bar{H}^{\otimes 2}(x)] + u_t \partial_t \nabla f_t(x)^\dagger \bar{H}(x) + \frac{1}{2}\partial_t^2 f_t(x).$$

is a continuous linear operator. The statement applies for spaces of families of functions as well (see Equation 26). Moreover, if $f \in G_\kappa^2([0,T]\times\mathbb{R}^d)$ with $f_t \in G_\kappa^\infty(\mathbb{R}^d)$, uniformly in t, then $\mathcal{F}f_t \in G_{\kappa+2}^\infty(\mathbb{R}^d)$, uniformly in t.

Proof The linearity of $\mathcal{F}$ is trivial. Now, given $f \in G^2_\kappa([0,T]\times\mathbb{R}^d)$ we have

$$\begin{aligned}\|\mathcal{F}f\|_{G_{\kappa+2}} \leq& \frac{9}{2}\|u\|^2_\infty \sum_{i,j}^d \|\partial_{i,j}f\|_{G_\kappa}\|\bar{H}_i\|_{G_1}\|\bar{H}_j\|_{G_1} \\ &+ 3\|u\|_\infty \sum_{i=1}^d \|\partial_t\partial_i f\|_{G_\kappa}\|\bar{H}_i\|_{G_1} + \frac{1}{2}\|\partial_t^2 f\|_{G_\kappa}\end{aligned}$$

From this we can see that $\|\mathcal{F}f\|_{G_{\kappa+2}} < \infty$, so $\mathcal{F}$ is well-defined. Furthermore, the bound on $\|\mathcal{F}f\|_{G_{\kappa+2}}$ is a scalar multiple of the norm on $G^2_\kappa([0,T]\times\mathbb{R}^d)$ proving the continuity. To show the last sentence note that $\|\partial^\alpha \mathcal{F}f\|_{G_{\kappa+2}}$ is bounded by a linear combination of the G_κ-norms of $f, \partial_t f, \partial_t^2 f$ and their derivatives, as well as $\|\bar{H}\|_{G_1}$ and the ∞-norms of the derivatives of $\bar{H}$. ■

Corollary 19 *There exists a function $C \in G(\mathbb{R}^d)$, such that*

$$|\varphi^h_t(x) - \varphi^0_t(x)| \leq hC(x),$$

for all $t \in [0,T], x \in \mathbb{R}^d$ and $h \in \mathcal{H}$. Consequently,

$$|\mathbb{E}\varphi^h_t(X^h_t) - \mathbb{E}\varphi^0_t(X^h_t)| = \mathcal{O}(h) \tag{37}$$

for all $t \in [0,T]$ and $h \in \mathcal{H}$.

Proof With δ defined as in Lemma 17 we have

$$\varphi^h - \varphi^0 = h\mathcal{F}\delta^h.$$

Now apply Lemma 17 and the fact that $\mathcal{F}$ maps into $G([0,T]\times\mathbb{R}^d)$. With this, Inequality (37) follows from Theorem 26 in the Appendix. ■

Lemma 20 *We have*

$$|\mathbb{E}\varphi^0_t(X^h_t) - \varphi^0_t(X^0_t)| = \mathcal{O}(h) \tag{38}$$

for all $t \in [0,T]$ and $h \in \mathcal{H}$.

Proof If we replace χ^h_k by $\tilde{X}^h_k$ in Lemma 12 and its extension in (30), then we can proceed with the proof in the same way to show

$$\mathbb{E}\varphi^0_{nh}(\tilde{X}^h_n) - \varphi^0_{nh}(X^0_{nh}) = h^2 \sum_{k=0}^{n-1} \mathbb{E}\Psi^h_{n,k}(\tilde{X}^h_k) + \mathcal{O}(h^2) \tag{39}$$

where

$$\Psi^h_{n,k}(x) := \mathcal{F}_{kh}(\mathbb{E}\varphi^0_{nh}(X^{h,\cdot}_{nh}))(x).$$

Here $X^{h,\cdot}_{nh}$ is a random field with variable initial value $x \in \mathbb{R}^d$.

We use the Itô-Taylor approximation in Lemma 11 to calculate $\mathbb{E}(\Delta\tilde{X}^h_n|\tilde{X}^h_n)$ and $\mathbb{E}((\Delta\tilde{X}^h_n)^{\otimes 2}|\tilde{X}^h_n)$, and estimate $\|\tilde{X}^h\|_{\#\beta}$ using Theorem 26.

Having established (39) next we consider the family

$$w^{h,r}_s(x) := \mathbb{E}\varphi^0_r(X^{h,s}_r(x)),$$

which satisfies $w \in G^\infty([0,T]\times\mathbb{R}^d)$, uniformly in h, r and s, by a straightforward extension of Remark 9. Therefore, Lemma 18 implies

$$|\Psi^h_{n,k}(x)| = |(\mathcal{F}_{kh}v^{h,nh})(x)| \leq C(1+|x|^\kappa),$$

for some $C > 0$ and $\kappa \in \mathbb{N}$. This proves (38) for $t = nh$.

Now consider an arbitrary $t \in [0,T]$ with $nh \leq t < (n+1)h$. Then Taylor's theorem, the Cauchy-Schwarz inequality and the fact that $(t-nh) \leq h$, imply

$$\begin{aligned}|\mathbb{E}\varphi^0_t(X^h_t) - \mathbb{E}\varphi^0_{nh}(X^h_{nh})| \leq& h|\mathbb{E}\partial_t\varphi^0_{nh}(X^h_{nh})| + \|\nabla\varphi^0_{nh}(X^h_{nh})\|_2\|\Delta\tilde{X}^h_n\|_2 \\ &+ \mathcal{O}(h^2),\end{aligned}$$

with some remainder $r \in G([0,T]\times\mathbb{R}^d)$. So,

$$|\mathbb{E}\varphi^0_t(X^h_t) - \mathbb{E}\varphi^0_{nh}(X^h_{nh})| = \mathcal{O}(h)$$

for all $h \in \mathcal{H}$ by Lemma 10, Theorem 26 and since $\nabla\varphi^0 \in G([0,T]\times\mathbb{R}^d)$ by the last statement of Lemma 18. Similarly,

$$|\varphi^0_t(X^0_t) - \varphi^0_{nh}(X^0_{nh})| \in \mathcal{O}(h),$$

for all $h \in \mathcal{H}$. Hence,

$$\begin{aligned}|\mathbb{E}\varphi^0_t(X^h_t) - \varphi^0_t(X^0_t)| \leq& |\mathbb{E}\varphi^0_t(X^h_t) - \mathbb{E}\varphi^0_{nh}(\tilde{X}^h_n)| \\ &+ |\mathbb{E}\varphi^0_{nh}(\tilde{X}^h_n) - \varphi^0_{nh}(X^0_{nh})| \\ &+ |\varphi^0_t(X^0_t) - \varphi^0_{nh}(X^0_{nh})| \\ \in& \mathcal{O}(h)\end{aligned}$$

for all $t \in [0,T]$ and $h \in \mathcal{H}$. ■

Proof of Theorem 2 Combining inequalities (37) and (38) gives us

$$\begin{aligned}|\mathbb{E}\varphi^h_t(X^h_t) - \varphi^0_t(X^0_t)| \leq& |\mathbb{E}\varphi^h_t(X^h_t) - \mathbb{E}\varphi^0_t(X^h_t)| + |\mathbb{E}\varphi^0_t(X^h_t) - \varphi^0_t(X^0_t)| \\ \in& \mathcal{O}(h)\end{aligned}$$

for all $h \in \mathcal{H}$. We conclude with the help of (34),

$$\mathbb{E}g(\chi^h_{T/h}) - \mathbb{E}g(X^h_T) = h\int_0^T \varphi^0_t(X^0_t)\,dt + \mathcal{O}(h^2).$$

■

5. Proof of Theorem 4

In this section our main aim is to prove Theorem 4, which extends the results from Theorems 1, 2 and 3. We have already used Theorem 4 to study the case of linear regression.

First, we need to introduce some additional facts and notation. Given tensors $B \in \mathbb{R}^{d\times l}$ and $C \in \mathbb{R}^{d\times k}$ we define their *outer product* $B \otimes C \in \mathbb{R}^{d\times(l+k)}$ by

$$(B\otimes C)_{i_1,\dots,i_l,j_1,\dots,j_k} = B_{i_1,\dots,i_l}C_{j_1,\dots,j_k},$$

and we set $B^{\otimes 2} := B \otimes B$ (as we did previously for vectors). Given vectors $u, v \in \mathbb{R}^d$ we also write $u \otimes v := uv^\dagger \in \mathbb{R}^{d\times d}$. Note that given a matrix A we have $u^\dagger A v = \langle A, u\otimes v\rangle$. Further, $\|v^{\otimes 2}\|_F \le |v|^2$.

In this section, to reduce notational clutter we again often omit the initial condition from the solution X of a given SDE and formulate statements for the mapping from the set of in initial conditions $\mathbb{R}^d$ to the collection of random variables $(X_t)_t$.

The next proposition gives alternative representations for the functions φ and $\nabla^2 v$ appearing in Theorems 1, 2 and 3.

Proposition 21 *Define*

$$\varphi_t^{\mathrm{id}} = (\varphi_t^1,\dots,\varphi_t^d) : \mathbb{R}^d \to \mathbb{R}^d,$$

where $\varphi_t^i := \varphi_t^g$ for $g(x) = x_i$ and $i \in \{1,\dots,d\}$. Then

$$\varphi_t^{\mathrm{id}} = \frac{1}{2}\langle \nabla^2 X_T^{0,t}, \bar{H}^{\otimes 2}\rangle - (\nabla X_T^{0,t})^\dagger \nabla \bar{H}(X_T^{0,t})^\dagger \bar{H} + \frac{1}{2}\nabla\bar{H}(X_T^{0,t})\bar{H}(X_T^{0,t}), \quad t\in[0,T].$$

Further, if $g \in G^\infty(\mathbb{R}^d)$, then

$$\varphi_t^g = \langle \nabla g(X_T^{0,t}), \varphi_t^{\mathrm{id}}\rangle,$$

and for any $S \in \mathbb{R}^{d\times d}$,

$$\langle \nabla^2 v_t^g, S\rangle = \langle \nabla^2 g(X_T^{0,t}), \nabla X_T^{0,t} S (\nabla X_T^{0,t})^\dagger\rangle + \langle \nabla g(X_T^{0,t}), \langle \nabla^2 X_T^{0,t}, S\rangle\rangle,$$

for all $t \in [0,T]$. Finally, if $\bar{H}$ is a conservative vector field and ∇X_t^0 is symmetric everywhere, then

$$\varphi_t^{\mathrm{id}} = \frac{1}{2}\langle \nabla^2 X_T^{0,t}, \bar{H}^{\otimes 2}\rangle - \frac{1}{2}\nabla\bar{H}(X_T^{0,t})\bar{H}(X_T^{0,t}), \quad t\in[0,T].$$

Remark 22 *Note that $\bar{H}$ is called conservative if it is the gradient of some function. For example, in case of SGD with replacement (see Example 1) we have $\bar{H} = \nabla\mathbb{E}[R_{\gamma(n)}]$ and so, indeed, $\bar{H}$ is conservative.*

On the other hand, characterizing the symmetry of ∇X_t^0 is a delicate issue that goes beyond the scope of this paper. A sufficient condition is that $\bar{H}$ is conservative and $\int_0^t \nabla\bar{H}(X_s)\,ds$ commutes with $\nabla\bar{H}(X_t^0)$, for all $t \ge 0$. In this case, we can solve the differential equation

$$d\nabla X_t^0 = \nabla\bar{H}(X_t^0)\nabla X_t^0\,dt, \quad \nabla X_0 = 1_{d\times d},$$

explicitly. The solution is then given by

$$\nabla X_t^0 = \exp\left(\int_0^t \nabla\bar{H}(X_s^0)\,ds\right), \quad t\in[0,T].$$

Since $\nabla H(X_s^0)$ is symmetric for all $s \in [0,T]$, so is ∇X_t^0. For more general conditions for the symmetry of ∇X_t^0 one may refer to Fetisov (2021).

Proof of Proposition 21 Recall the definition of φ in Equation (10). Note that $X_T^{0,t} = X_{T-t}^0$. Thus,

$$\begin{aligned}
\partial_t X_{T-t}^0 =& -\partial_t X_{T-t}^0 = -\bar{H}(X_{T-t}^0),\\
\partial_t^2 X_{T-t}^0 =& -\partial_t(\bar{H}(X_{T-t}^0)) = \nabla\bar{H}(X_{T-t}^0)\bar{H}(X_{T-t}^0),\\
\partial_t \nabla X_{T-t}^0 =& -\nabla\bar{H}(X_{T-t}^0)\nabla X_{T-t}^0,
\end{aligned} \tag{40}$$

where the last equation follows from Theorem 29. Moreover,

$$\partial_t(\bar{H}(X_T^{0,t})) = \nabla\bar{H}(X_T^{0,t})\partial_t X_T^{0,t} = -\nabla\bar{H}(X_T^{0,t})\bar{H}(X_T^{0,t}),$$

and the solution to this linear equation can be expressed in terms of the solution to associated matrix equation (40)

$$\bar{H}(X_T^{0,t}) = \nabla X_T^{0,t}\bar{H}, \quad t\in[0,T]. \tag{41}$$

For all $g\in G^\infty$, we have

$$\begin{aligned}
\partial_t v_t^g =& \sum_i \partial_i g(X_T^{0,t})(\partial_t X_T^{0,t})^i,\\
\partial_t^2 v_t^g =& \sum_{i,j}(\partial_{ij} g(X_T^{0,t})(\partial_t X_T^{0,t})^i(\partial_t X_T^{0,t})^j + \partial_i g(X_T^{0,t})(\partial_t^2 X_T^{0,t})^i)\\
\partial_k v_t^g =& \sum_i \partial_i g(X_T^{0,t})(\partial_k X_T^{0,t})^i\\
\partial_k\partial_t v_t^g =& \sum_{i,j}(\partial_{ij} g(X_T^{0,t})(\partial_t X_T^{0,t})^j(\partial_k X_T^{0,t})^i + \partial_i g(X_T^{0,t})(\partial_k\partial_t X_T^{0,t})^i)\\
\partial_{kl} v_t^g =& \sum_{i,j}(\partial_{ij} g(X_T^{0,t})(\partial_l X_T^{0,t})^j(\partial_k X_T^{0,t})^i + \partial_i g(X_T^{0,t})(\partial_{kl} X_T^{0,t})^i).
\end{aligned}$$

Thus,

$$\begin{aligned}
\partial_t^2 v_t^g =& \langle\nabla^2 g(X_T^{0,t}), (\partial_t X_T^{0,t})^{\otimes 2}\rangle + \langle\nabla g(X_T^{0,t}), \partial_t^2 X_T^{0,t}\rangle,\\
\langle\partial_t\nabla v_t^g, \bar{H}\rangle =& \langle\nabla^2 g(X_T^{0,t}), \nabla X_T^{0,t}\bar{H}\otimes\partial_t X_T^{0,t}\rangle + \langle\nabla g(X_T^{0,t}), \partial_t(\nabla X_T^{0,t})^\dagger\bar{H}\rangle.
\end{aligned}$$

Moreover, note that for matrices $A, B\in\mathbb{R}^{d\times d}$ we have $\langle A, B\rangle = \mathrm{tr}(A^\dagger B)$. In particular, for $A, B, C, S\in\mathbb{R}^{d\times d}$, using the cyclic property of the matrix trace,

$$\langle ABC, S\rangle = \mathrm{tr}(C^\dagger B^\dagger A^\dagger S) = \mathrm{tr}(B^\dagger A^\dagger S C^\dagger) = \langle B, A^\dagger S C^\dagger\rangle.$$

Thus,

$$\begin{aligned}
\langle\nabla^2 v_t^g, S\rangle =& \langle(\nabla X_T^{0,t})^\dagger\nabla^2 g(X_T^{0,t})\nabla X_T^{0,t}, S\rangle + \langle\nabla g(X_T^{0,t}), \langle\nabla^2 X_T^{0,t}, S\rangle\rangle\\
=& \langle\nabla^2 g(X_T^{0,t}), \nabla X_T^{0,t} S(\nabla X_T^{0,t})^\dagger\rangle + \langle\nabla g(X_T^{0,t}), \langle\nabla^2 X_T^{0,t}, S\rangle\rangle, \quad S\in\mathbb{R}^{d\times d}.
\end{aligned}$$

Further, note that by (41)

$$\nabla X_T^{0,t}\bar{H}^{\otimes 2}(\nabla X_T^{0,t})^\dagger = \nabla X_T^{0,t}\bar{H}\bar{H}^\dagger(\nabla X_T^{0,t})^\dagger = \bar{H}(X_T^{0,t})\bar{H}(X_T^{0,t})^\dagger = \bar{H}(X_T^{0,t})^{\otimes 2},$$

and so

$$\begin{aligned}&(\partial_t X_T^{0,t})^{\otimes 2} + 2\nabla X_T^{0,t}\bar{H} \otimes \partial_t X_T^{0,t} + \nabla X_T^{0,t}\bar{H}^{\otimes 2}(\nabla X_T^{0,t})^\dagger \\ =&\bar{H}(X_{T-t}^0)^{\otimes 2} - 2\bar{H}(X_T^{0,t})^{\otimes 2} + \bar{H}(X_{T-t}^0)^{\otimes 2} \\ =&0.\end{aligned}$$

Thus,

$$\begin{aligned}\varphi_t^g =&\frac{1}{2}\langle \nabla^2 v_t^g, \bar{H}^{\otimes 2}\rangle + \langle \partial_t \nabla v_t^g, \bar{H}\rangle + \frac{1}{2}\partial_t^2 v_t^g \\ =&\langle \nabla g(X_T^{0,t}), \frac{1}{2}\langle \nabla^2 X_T^{0,t}, \bar{H}^{\otimes 2}\rangle + \partial_t(\nabla X_T^{0,t})^\dagger \bar{H} + \frac{1}{2}\partial_t^2 X_T^{0,t}\rangle.\end{aligned}$$

For $g(x) = x_i$ we have $\nabla g(x) = e_i$, and therefore

$$\varphi_t^{\mathrm{id}} = \frac{1}{2}\langle \nabla^2 X_T^{0,t}, \bar{H}^{\otimes 2}\rangle + \partial_t(\nabla X_T^{0,t})^\dagger \bar{H} + \frac{1}{2}\partial_t^2 X_T^{0,t}, \quad t \in [0,T].$$

Moreover, using once more the equations (40), we may rewrite φ^{id} as follows

$$\varphi_t^{\mathrm{id}} = \frac{1}{2}\langle \nabla^2 X_T^{0,t}, \bar{H}^{\otimes 2}\rangle - (\nabla X_T^{0,t})^\dagger \nabla \bar{H}(X_T^{0,t})^\dagger \bar{H} + \frac{1}{2}\nabla \bar{H}(X_T^{0,t})\bar{H}(X_T^{0,t}).$$

Furthermore, for an arbitrary $g \in G^\infty$ we have

$$\varphi_t^g = \langle \nabla g(X_T^{0,t}), \varphi_t^{\mathrm{id}}\rangle, \quad t \in [0,T].$$

Finally, suppose $\bar{H}$ is conservative and ∇X_t^0 is symmetric everywhere. Then $\nabla \bar{H}(X_t^0)$ is symmetric as well for all $t \in [0,T]$, since higher partial derivatives commute. Hence,

$$\nabla X_t^0 \nabla \bar{H}(X_t^0) = (\partial_t \nabla X_t^0)^\dagger = \partial_t \nabla X_t^0 = \bar{H}(X_t^0)\nabla X_t^0, \quad t \in [0,T].$$

Consequently, by (41),

$$(\nabla X_T^{0,t})^\dagger \nabla \bar{H}(X_T^{0,t})\bar{H} = \nabla \bar{H}(X_T^{0,t})\nabla X_T^{0,t}\bar{H} = \nabla \bar{H}(X_T^{0,t})\bar{H}(X_T^{0,t}),$$

and so φ^{id} simplifies, as desired. ■

For the remainder of the section we consider an objective function $\mathcal{R} \in C^\infty(\mathbb{R}^d) \cap G_2(\mathbb{R}^d)$ with $\nabla \mathcal{R} \in \mathrm{Lip}^\infty(\mathbb{R}^d)$ and set $g = \mathcal{R}$. Since we are looking at a minimization problem the objective function is arguably the most important test function g to consider. As we can see in Theorem 4, the linear error terms for the various continuous-time approximations of SGD have a particularly nice form in this case.

Proof of Theorem 4 By the last equation in Proposition 21, we have

$$\varphi_t^{\mathrm{id}}(X_t^0) = \frac{1}{2}\langle \nabla^2 X_T^{0,t}(X_t^0), \nabla \mathcal{R}(X_t^0)^{\otimes 2}\rangle - \frac{1}{2}\nabla^2 \mathcal{R}(X_T^0)\nabla \mathcal{R}(X_T^0),$$

and further

$$\varphi_t^{\mathcal{R}}(X_t^0) = -\frac{1}{2}\langle \nabla^2 \mathcal{R}(X_T^0), \nabla \mathcal{R}(X_T^0)^{\otimes 2}\rangle + \frac{1}{2}\langle \nabla \mathcal{R}(X_T^0), \langle \nabla^2 X_T^{0,t}(X_t^0), \nabla \mathcal{R}(X_t^0)^{\otimes 2}\rangle\rangle.$$

Hence, by Theorem 2,

$$\begin{aligned}2\,\mathrm{LE}(X^{\mathrm{NCC}}) &= \int_0^T \varphi_t^{\mathcal{R}}(X_t^0)\,dt \\ &= -T\langle \nabla^2\mathcal{R}(X_T^0), \nabla\mathcal{R}(X_T^0)^{\otimes 2}\rangle \\ &\quad + \int_0^T \langle \nabla\mathcal{R}(X_T^0), \langle \nabla^2 X_T^{0,t}(X_t^0), \nabla\mathcal{R}(X_t^0)^{\otimes 2}\rangle\rangle\,dt.\end{aligned}$$

Moreover, by Theorem 1 and 3, as well Proposition 21,

$$\begin{aligned}2\,\mathrm{LE}(X) =& 2\,\mathrm{LE}(X^{\mathrm{NCC}}) \\ &+ \int_0^T \langle \nabla^2\mathcal{R}(X_T^0), \nabla X_T^{0,t}(X_t^0)\tilde{D}(X_t^0)\nabla X_T^{0,t}(X_t^0)\rangle\,dt \\ &+ \int_0^T \langle \nabla\mathcal{R}(X_T^0), \langle \nabla^2 X_T^{0,t}(X_t^0), \tilde{D}(X_t^0)\rangle\rangle\,dt \\ =& \langle \nabla^2\mathcal{R}(X_T^0), \alpha_T - T\nabla\mathcal{R}(X_T^0)^{\otimes 2}\rangle + \langle \nabla\mathcal{R}(X_T^0), \beta_T\rangle,\end{aligned}$$

where $\tilde{D} = \Sigma - D$. To prove (17), we estimate using Theorem 29

$$\begin{aligned}\|\alpha_T\|_F \leq& \int_0^T \|\nabla X_T^{0,t}(X_t^0)\|_F^2 \|\Sigma(X_t^0) - D(X_T^0)\|_F\,dt \\ \leq& d\int_0^T \frac{M_T^2}{M_t^2}\|\Sigma(X_t^0) - D(X_t^0)\|_F\,dt = d\xi_T^D.\end{aligned}$$

Further, note that in our case (41) takes the form

$$\nabla\mathcal{R}(X_t^0) = \nabla X_t^0 \nabla\mathcal{R}, \quad t \geq 0,$$

and so

$$|\nabla\mathcal{R}(X_t^0)| \leq \|\nabla X_t^0\|_F |\nabla\mathcal{R}| \leq \sqrt{d} M_t |\nabla\mathcal{R}|, \quad t \geq 0.$$

Equation (45) in Theorem 29 implies

$$\|\nabla^2 X_T^{0,t}(X_t^0)\|_F \leq d\frac{M_T}{M_t}\int_t^T \frac{M_s}{M_t}\|\nabla^3\mathcal{R}(X_s^0)\|_F\,ds.$$

Note that $\|v^{\times 2}\|_F \leq |v|^2, v \in \mathbb{R}^d$. Further, for tensors A and B we have the following generalization of the Cauchy-Schwarz inequality

$$\|\langle A, B\rangle\|_F \leq \|A\|_F \|B\|_F.$$

Therefore,

$$\begin{aligned}|\beta_T| \leq& \int_0^T \|\nabla^2 X_T^{0,t}(X_t^0)\|_F (|\nabla\mathcal{R}(X_t^0)|^2 + \|\Sigma(X_t^0) - D(X_t^0)\|_F)\,dt \\ \leq& d\int_0^T \frac{M_T}{M_t^2}|\nabla\mathcal{R}(X_t^0)|^2 \left(\int_t^T M_s\|\nabla^3\mathcal{R}(X_s^0)\|_F\,ds\right)\,dt \\ &+ d\int_0^T \frac{M_T}{M_t^2}\|\Sigma(X_t^0) - D(X_t^0)\|_F \left(\int_t^T M_s\|\nabla^3\mathcal{R}(X_s^0)\|_F\,ds\right)\,dt \\ \leq& d^2 T M_T |\nabla\mathcal{R}|^2 \zeta_T + d M_T^{-1}\xi_T^D \zeta_T.\end{aligned}$$

Thus, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
2|\operatorname{LE}(X)| \leq & |\langle \nabla^2 \mathcal{R}(X_T^0), \alpha_T - T\nabla\mathcal{R}(X_T^0)^{\otimes 2}\rangle| + |\langle \nabla\mathcal{R}(X_T^0), \beta_T\rangle| \\
\leq & T\|\nabla^2\mathcal{R}(X_T^0)\|_F |\nabla\mathcal{R}(X_T^0)|^2 + \|\nabla^2\mathcal{R}(X_T^0)\|_F\|\alpha_T\|_F \\
& + |\nabla\mathcal{R}(X_T^0)||\beta_T| \\
\leq & dTM_T^2|\nabla\mathcal{R}|^2\|\nabla^2\mathcal{R}(X_T^0)\|_F + d\|\nabla^2\mathcal{R}(X_T^0)\|_F\xi_T^D \\
& + \sqrt{d}M_T|\nabla\mathcal{R}|(d^2TM_T|\nabla\mathcal{R}|^2\zeta_T + dM_T^{-1}\xi_T^D\zeta_T) \\
= & dTM_T^2|\nabla\mathcal{R}|^2(\|\nabla^2\mathcal{R}(X_T^0)\|_F + d^{3/2}|\nabla\mathcal{R}|\zeta_T) \\
& + d\xi_T^D(\|\nabla^2\mathcal{R}(X_T^0)\|_F + \sqrt{d}|\nabla\mathcal{R}|\zeta_T).
\end{aligned}$$

■

6. Derivations and Proofs for Section 3

In this section we give proper justifications for the results of Section 3.

6.1 Quadratic Objectives

Here, we derive the linear error terms for the three continuous-time approximations when the objective function is quadratic. This includes ordinary linear regression with SGD using the population risk, but the derivation applies more generally.

Suppose we are given a symmetric and positive definite matrix $\kappa \in \mathbb{R}^{d\times d}$ and a quadratic form

$$\mathcal{R}(\theta) = \theta^\dagger\kappa\theta + \theta^\dagger c' + d', \quad \theta \in \mathbb{R}^d,$$

where $c' \in \mathbb{R}^d$ and $d' \in \mathbb{R}$. Then $\mathcal{R}$ has a global minimum $\theta^* \in \mathbb{R}^d$ and so we may rewrite it as

$$\mathcal{R}(\theta) = \langle \kappa, (\theta - \theta^*)^{\otimes 2}\rangle + d, \quad \theta \in \mathbb{R}^d$$

for some $d \in \mathbb{R}$. Now, consider SGD with $\bar{H}(\theta) = -\nabla\mathcal{R}(\theta)$. The gradient flow equation

$$dX_t^0 = -\nabla\mathcal{R}(X_t^0) = -\kappa(X_t^0 - \theta^*)\,dt,$$

has the unique solution

$$X_t^0(\theta) = e^{-t\kappa}(\theta - \theta^*) + \theta^*, \quad t \in [0, T],$$

for every initial condition $\theta \in \mathbb{R}^d$. Note that $X_t^0(\theta) \to X_\infty^0(\theta) = \theta^*$, as $t \to \infty$, for every $\theta \in \mathbb{R}^d$. Further, $\nabla^2\mathcal{R} = \kappa, \nabla^3\mathcal{R} = 0, \nabla X_t^0 = e^{-t\kappa}, \nabla^2 X^0 = 0, \nabla X_T^{0,t}(X_t^0) = e^{-(T-t)\kappa}, \nabla\mathcal{R}(X_T^0(\theta)) = -e^{-T\kappa}\kappa(\theta - \theta^*)$, and so

$$\begin{aligned}
\langle \kappa, \nabla\mathcal{R}(X_T^0(\theta))^{\otimes 2}\rangle = & \operatorname{tr}(\kappa^\dagger e^{-T\kappa}\kappa(\theta-\theta^*)(\theta-\theta^*)^\dagger\kappa^\dagger e^{-T\kappa^\dagger}) \\
= & \langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle.
\end{aligned}$$

Now, consider $D \in \{0, \Sigma(\theta^*), \Sigma\}$ and

$$dX_t^h = -\nabla \mathcal{R}(X_t^h) + \sqrt{hD(X_t^h)}\, dW_t, \quad t \in [0,T], h \in \mathcal{H}.$$

By Theorem 4 we have

$$\mathrm{LE}(X) = -\frac{1}{2} T \langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2} \rangle + \frac{1}{2} \langle \kappa, \alpha_T^D \rangle, \tag{42}$$

where

$$\alpha_T^D = \int_0^T e^{-(T-t)\kappa} (\Sigma(X_t^0) - D(X_t^0)) e^{-(T-t)\kappa}\, dt.$$

6.2 Derivation of the Covariance Matrix of the Gradient Noise

Recall the statistical learning setting in Section 3.1. Then for $S(\theta) := \mathrm{Cov}[\nabla_\theta \mathcal{R}_{\boldsymbol{x},\boldsymbol{y}}(\theta)]$ we have

$$\begin{aligned}
S(\theta) =& \mathbb{E}[(\langle \theta, \boldsymbol{x} \rangle - \boldsymbol{y})^2 \boldsymbol{x}^{\otimes 2}] - (\kappa(\theta - \theta^*))^{\otimes 2} \\
=& \mathbb{E}[(\langle \theta - \theta^*, \boldsymbol{x} \rangle - \boldsymbol{\varepsilon})^2 \boldsymbol{x}^{\otimes 2}] - \kappa(\theta - \theta^*)^{\otimes 2} \kappa^\dagger \\
=& \mathbb{E}[\langle \theta - \theta^*, \boldsymbol{x} \rangle^2 \boldsymbol{x}^{\otimes 2}] - 2\mathbb{E}[\varepsilon \langle \theta - \theta^*, \boldsymbol{x} \rangle \boldsymbol{x}^{\otimes 2}] \\
& + \mathbb{E}[\boldsymbol{\varepsilon}^2 \boldsymbol{x}^{\otimes 2}] - \kappa(\theta - \theta^*)^{\otimes 2} \kappa^\dagger \\
=& \langle \mu_x^4, (\theta - \theta^*)^{\otimes 2} \rangle - \kappa(\theta - \theta^*)^{\otimes 2} \kappa^\dagger + \sigma_\varepsilon^2 \kappa \\
=& \langle \mu_x^4 - \kappa^{\otimes 2}, (\theta - \theta^*)^{\otimes 2} \rangle + \sigma_\varepsilon^2 \kappa
\end{aligned}$$

Recall Example 2 (a). We will now derive the explicit formula for S given there. Let τ be a permutation of the set $\{1, \dots, l\}$ and $B \in \mathbb{R}^{d \times l}$ an l-tensor. Then we write $B_\tau \in \mathbb{R}^{d \times l}$ for

$$(B_\tau)_{i_1, \dots, i_l} = B_{i_{\tau(1)}, \dots, i_{\tau(l)}}.$$

For example if B is matrix, then $B^\dagger = B_{(12)}$. Here we use the cycle notation for permutations. By Isserli's theorem (see Bose, 2021), the joint fourth moments of a centered Gaussian satisfy

$$\mu_x^4 = \kappa^{\otimes 2} + \kappa_{(23)}^{\otimes 2} + \kappa_{(13)}^{\otimes 2}.$$

Given matrices $U, A \in \mathbb{R}^{d \times d}$ we have

$$\begin{aligned}
\langle U_{(23)}^{\otimes 2}, A \rangle_{i,j} =& \sum_{k,l} U_{i,k} U_{j,l} A_{k,l} \\
=& U A U^\dagger, \\
\langle U_{(13)}^{\otimes 2}, A \rangle_{i,j} =& \sum_{k,l} U_{k,j} U_{i,l} A_{k,l} \\
=& U A^\dagger U.
\end{aligned}$$

Therefore, $S(\theta) = 2\kappa(\theta - \theta^*)^{\otimes 2} \kappa + \sigma_\varepsilon^2 \kappa$.

Proof of Proposition 5 Recall Equation (42). The first equation in Proposition 5 follows by setting $\alpha_T = 0$ in the linear error expansion. Moreover,

$$\Sigma(X_t^0) - \Sigma(\theta^*) = 2\frac{B^{\mathrm{Eq}}}{B}\kappa e^{-t\kappa}(\theta-\theta^*)^{\otimes 2}e^{-t\kappa}\kappa,$$

and so

$$\begin{aligned}\alpha_T^{\Sigma(\theta^*)} =& 2\frac{B^{\mathrm{Eq}}}{B}\int_0^T e^{-(T-t)\kappa}\kappa e^{-t\kappa}(\theta-\theta^*)^{\otimes 2}e^{-t\kappa}\kappa e^{-(T-t)\kappa}\,dt\\ =& 2T\frac{B^{\mathrm{Eq}}}{B}(\kappa e^{-T\kappa}(\theta-\theta^*))^{\otimes 2}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{LE}(X^{\mathrm{CC}}) =& \mathrm{LE}(X^{\mathrm{NCC}}) + T\frac{B^{\mathrm{Eq}}}{B}\langle\kappa, (\kappa e^{-T\kappa}(\theta-\theta^*))^{\otimes 2}\rangle\\ =& T\left(\frac{B^{\mathrm{Eq}}}{B}-\frac{1}{2}\right)\langle\kappa^3 e^{-2T\kappa}, (\theta-\theta^*)^{\otimes 2}\rangle.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathrm{LE}(X^0) =& \mathrm{LE}(X^{\mathrm{CC}}) + \frac{1}{2}\langle\kappa, \int_0^T e^{-(T-t)\kappa}\Sigma(\theta^*)e^{-(T-t)\kappa}\rangle\,dt\\ =& \mathrm{LE}(X^{\mathrm{CC}}) + \frac{1}{2B}\sigma_\varepsilon^2\langle\kappa^2, \int_0^T e^{-2(T-t)\kappa}\,dt\rangle.\end{aligned}$$

Finally, since κ is positive definite, we may simplify

$$\frac{1}{2B}\sigma_\varepsilon^2\langle\kappa^2, \int_0^T e^{-2(T-t)\kappa}\,dt\rangle = \frac{1}{4B}\sigma_\varepsilon^2\langle\kappa^2, (1_{d\times d} - e^{-2\kappa T})\kappa^{-1}\rangle = \frac{1}{4B}\sigma_\varepsilon^2\langle\kappa, 1_{d\times d} - e^{-2\kappa T}\rangle.$$

■

The following lemma is used in the proof of Theorem 6.

Lemma 23 *Let $a, b_1, b_2, B > 0$ with $b_1 < b_2$ and set $e_i = -a + \frac{b_i}{B}$ Then,*

$$\mathrm{sgn}(|e_1| - |e_2|) = \mathrm{sgn}\left(B - \frac{b_1+b_2}{2a}\right).$$

Proof Note that $B \le \frac{b_i}{a}$ if and only if $e_i \ge 0$, and $B \ge \frac{b_i}{a}$ if and only if $e_i \le 0$. Moreover,

$$\frac{b_1}{a} < \frac{b_1+b_2}{2a} < \frac{b_2}{a}.$$

Thus, we have $|e_1| < |e_2|$ if and only if

(a) $B \le \frac{b_1}{a}$ and $e_1 < e_2$, or

(b) $\frac{b_2}{a} < B \le \frac{b_1}{a}$ and $e_1 < -e_2$, or

(c) $\frac{b_1}{a} < B \leq \frac{b_2}{a}$ and $-e_1 < e_2$, or

(d) $B > \frac{b_2}{a}$ and $-e_1 < -e_2$.

Since $e_1 < e_2$, case (d) can never occur and (a) is equivalent to $B \leq \frac{b_1}{a}$. Further, since $b_1 < b_2$, (b) is also impossible. Moreover, (c) is equivalent to

$$\frac{b_1}{a} \leq B < \frac{b_1 + b_2}{2a}.$$

Putting (a) and (c) together yields

$$|e_1| < |e_2| \Leftrightarrow B < \frac{b_1 + b_2}{2a}.$$

Finally, since

$$|e_1| = |e_2| \Leftrightarrow e_1 = -e_2 \Leftrightarrow B = \frac{b_1 + b_2}{2a},$$

the result follows. ■

Proof of Theorem 6 Set

$$a := \frac{1}{2}T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle, b := B^{\mathrm{Eq}} T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle, c := \frac{1}{4}\sigma_\varepsilon^2 \langle \kappa, 1_{d\times d} - e^{-2\kappa T}\rangle > 0.$$

By definition and Proposition 5

$$\mathrm{LE}(X^{\mathrm{NCC}}) = -a, \quad \mathrm{LE}(X^{\mathrm{CC}}) = -a + \frac{b}{B}, \quad \mathrm{LE}(X^0) = -a + \frac{b}{B} + \frac{c}{B}.$$

Lemma 23 implies

$$\begin{aligned}
B < \frac{b}{2a} &\Leftrightarrow |\,\mathrm{LE}(X^{\mathrm{NCC}})| < |\,\mathrm{LE}(X^{\mathrm{CC}})|, \quad B > \frac{b}{2a} \Leftrightarrow |\,\mathrm{LE}(X^{\mathrm{NCC}})| > |\,\mathrm{LE}(X^{\mathrm{CC}})|,\\
B < \frac{b+c}{2a} &\Leftrightarrow |\,\mathrm{LE}(X^{\mathrm{NCC}})| < |\,\mathrm{LE}(X^0)|, \quad B > \frac{b+c}{2a} \Leftrightarrow |\,\mathrm{LE}(X^{\mathrm{NCC}})| > |\,\mathrm{LE}(X^0)|,\\
B < \frac{2b+c}{2a} &\Leftrightarrow |\,\mathrm{LE}(X^{\mathrm{CC}})| < |\,\mathrm{LE}(X^0)|, \quad B > \frac{2b+c}{2a} \Leftrightarrow |\,\mathrm{LE}(X^0)| > |\,\mathrm{LE}(X^{\mathrm{CC}})|.
\end{aligned}$$

Further,

$$B^{\mathrm{Eq}} = \frac{b}{2a}, \quad \frac{c}{2a} = \frac{\sigma_\varepsilon^2\langle \kappa, 1 - e^{-2T\kappa}\rangle}{4T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle}, \quad B^{\mathrm{GF}} = \frac{2b+c}{2a},$$

and so the cases (i) - (iv) are proven. Finally,

$$\begin{aligned}
&\mathrm{LE}(X^{\mathrm{CC}}) = 0 \Leftrightarrow B = \frac{b}{a} = 2B^{\mathrm{Eq}},\\
&\mathrm{LE}(X^{\mathrm{CC}}) = 0 \Leftrightarrow B = \frac{b+c}{a} = \frac{2b+c}{2a} + \frac{c}{2a} = B^{\mathrm{GF}} + B^{\mathrm{GF}} - 2B^{\mathrm{Eq}},
\end{aligned}$$

proving (v) and (vi). ■

Remark 24 *There are few additional statements one can make, adding to the list in Theorem 6. Firstly,*

(i) $X^0 \asymp X^{\mathrm{NCC}}$, *if* $B = B^{\mathrm{GF}} - B^{\mathrm{Eq}}$,

(ii) $X^0 \asymp X^{\mathrm{CC}}$, *if* $B = B^{\mathrm{GF}}$,

(iii) $\mathrm{LE}(X^0) = 0$, *if* $B = 2(B^{\mathrm{GF}} - B^{\mathrm{Eq}})$.

Note however that these will almost never occur in practice because it is unlikely that B^{GF} *is an integer. That is, unless one specifically designs the problem in such a way. On the other hand, notice that* $B^{\mathrm{Eq}} = 1$ *if* $\boldsymbol{x}$ *is Gaussian and* $B^{\mathrm{Eq}} = 4$ *if* $d = 1$ *and* $\boldsymbol{x}$ *is exponentially distributed and so the case (ii) in Theorem 6 can realistically occur in applications.*

Moreover, note that for $B^{\mathrm{Eq}} = 0$ *we have* $\Sigma(\theta) = \Sigma(\theta^*)$ *for all* $\theta \in \mathbb{R}^d$ *and so* $X^{\mathrm{CC}} = X^{\mathrm{NCC}}$. *In particular, this happens for* $d = 1$ *and if* $\boldsymbol{x}$ *has a symmetric Rademacher distribution, since then* $\operatorname{Kurt} x = 1$ *(recall example 2). Thus, we are left with the cases*

(i) $X^0 \prec X^{\mathrm{NCC}}$, *if* $B < B^{\mathrm{GF}}$,

(ii) $X^{\mathrm{NCC}} \prec X^0$, *if* $B > B^{\mathrm{GF}}$.

6.3 Explicit Formulas for the Expected Risk of the Continuous-Time Approximations of SGD for Linear Regression

Here, we derive explicit formulas for the expected (excess) population risk for four continuous-time approximation of SGD for linear regression. These are used in the numerical experiments to compute the continuous-time half of the weak error. Firstly, recall the following stochastic differential equations from Section 3.2

$$\begin{aligned} dX_t^0 &= -\kappa(X_t^0 - \theta^*)\,dt, \\ dX_t^{\mathrm{NCC},h} &= -\kappa(X_t^h - \theta^*)\,dt + \sqrt{\frac{h}{B}}\sqrt{2B^{\mathrm{Eq}}\kappa(X_t^{\mathrm{NCC},h} - \theta^*)^{\otimes 2}\kappa + \sigma_\varepsilon^2\kappa}\,dW_t, \\ dX_t^{\mathrm{CC},h} &= -\kappa(X_t^{\mathrm{CC},h} - \theta^*)\,dt + \sqrt{\frac{h}{B}\sigma_\varepsilon^2\kappa}\,dW_t. \end{aligned}$$

We also consider the following second-order diffusion approximation of SGD

$$dX_t^{2,h} = -\kappa\left(1_{d\times d} + \kappa\frac{h}{2}\right)(X_t^h - \theta^*)\,dt + \sqrt{\frac{h}{B}}\sqrt{2B^{\mathrm{Eq}}\kappa(\theta - \theta^*)^{\otimes 2}\kappa + \sigma_\varepsilon^2\kappa}\,dW_t.$$

For simplicity we set $d = 1$ and so $B^{\mathrm{Eq}} = \frac{1}{2}(\operatorname{Kurt}\boldsymbol{x} - 1)$. The next Proposition gives explicit formulas for the expected *excess population risk* $\mathbb{E}[\mathcal{R}^e(Y_t)]$ for $Y \in \{X^0, X^{\mathrm{NCC},h}, X^{\mathrm{CC},h}, X^{2,h}\}$, where $\mathcal{R}^e(\theta) = \frac{1}{2}(\theta - \theta^*)^2$. The actual population risk is also given by $\mathcal{R} = \kappa\mathcal{R}^e + \frac{\sigma_\varepsilon^2}{2}$. Note that

$$\mathcal{R}^e(\theta) - \mathcal{R}^e(\tilde{\theta}) = \frac{1}{\kappa}(\mathcal{R}(\theta) - \mathcal{R}(\tilde{\theta})), \quad \theta, \tilde{\theta} \in \mathbb{R}.$$

Proposition 25 *Define*

$$\zeta^h = 1 - \frac{h}{2B}\kappa(\mathrm{Kurt}\,\boldsymbol{x} - 1), \quad \xi^h := \zeta^h + \frac{h}{2}\kappa = 1 + \frac{h}{2B}\kappa(B + 1 - \mathrm{Kurt}\,\boldsymbol{x}), \quad h \in [0,1).$$

Then, we have

$$\begin{aligned}
\mathcal{R}^e(X_t^0) =& e^{-2\kappa t}\mathcal{R}^e(\theta),\\
\mathbb{E}[\mathcal{R}^e(X_t^{\mathrm{CC},h})] =& e^{-2\kappa t}\mathcal{R}^e(\theta) + \frac{h\sigma_\varepsilon^2}{4B}(1 - e^{-2\kappa t}),\\
\mathbb{E}[\mathcal{R}^e(X_t^{\mathrm{NCC},h})] =& e^{-2\kappa\zeta_h t}\mathcal{R}^e(\theta) + \frac{h\sigma_\varepsilon^2}{4B\zeta_h}(1 - e^{-2\kappa\zeta_h t}),\\
\mathbb{E}[\mathcal{R}^e(X_t^{2,h})] =& e^{-2\kappa\xi_h t}\mathcal{R}^e(\theta) + \frac{h\sigma_\varepsilon^2}{4B\xi_h}(1 - e^{-2\kappa\xi_h t}),
\end{aligned}$$

for all $h \in (0,1)$ *and* $t \geq 0$.

Proof Recall that

$$X_t^0 = e^{-\kappa t}(\theta - \theta^*) + \theta^*,$$

and so

$$\mathcal{R}^e(X_t^0) = e^{-2\kappa t}\mathcal{R}^e(\theta).$$

Further, $X^{\mathrm{CC},h}$ is an Ornstein-Uhlenbeck process and so

$$X_t^{\mathrm{CC},h} = X_t^0 + \sqrt{\frac{h\sigma_\varepsilon^2}{2B}}W_{1-e^{-2\kappa t}}.$$

Hence,

$$\mathbb{E}[\mathcal{R}^e(X_t^{\mathrm{CC},h})] = e^{-2\kappa t}\mathcal{R}^e(\theta) + \frac{h\sigma_\varepsilon^2}{4B}(1 - e^{-2\kappa t}).$$

Now, by Itô's formula

$$\begin{aligned}
d\mathcal{R}^e(X_t^{\mathrm{NCC},h}) =& -\kappa(X_t^{\mathrm{NCC},h} - \theta^*)^2 + \frac{h}{2B}\kappa^2(\mathrm{Kurt}\,\boldsymbol{x} - 1)(X_t^{\mathrm{NCC},h} - \theta^*)^2 + \frac{h}{2B}\kappa\sigma_\varepsilon^2\,dt + M_t\\
=& \left(\frac{h}{B}\kappa^2(\mathrm{Kurt}\,\boldsymbol{x} - 1) - 2\kappa\right)\mathcal{R}^e(X_t^{\mathrm{NCC},h}) + \frac{h}{2B}\kappa\sigma_\varepsilon^2\,dt + M_t
\end{aligned}$$

where M is a martingale starting in 0, a.s. Hence, by optional stopping

$$d\mathbb{E}[\mathcal{R}^e(X_t^{\mathrm{NCC},h})] = -2\kappa\zeta_h\mathbb{E}[\mathcal{R}^e(X_t^{\mathrm{NCC},h})] + \frac{h}{2B}\kappa\sigma_\varepsilon^2\,dt,$$

and so

$$\mathbb{E}[\mathcal{R}^e(X_t^{\mathrm{NCC},h})] = e^{-2\kappa\zeta_h t}\mathcal{R}^e(\theta) + \frac{h\sigma_\varepsilon^2}{4B\zeta_h}(1 - e^{-2\kappa\zeta_h t}).$$

Similarly,

$$\mathbb{E}[\mathcal{R}^e(X_t^{2,h})] = e^{-2\kappa\xi_h t}\mathcal{R}^e(\theta) + \frac{h\sigma_\varepsilon^2}{4B\xi_h}(1 - e^{-2\kappa\xi_h t}).$$

■

Acknowledgments

The authors would like to express their gratitude to the three anonymous reviewers of this article. Their detailed assessments and ideas for extensions have helped the authors to improve the first draft of this paper significantly.

Appendix A. A Remark on Kurtosis

The *kurtosis* of a distribution is its standardized fourth central moment. That is, given a random variable Z with $\mathbb{E}Z^4 < \infty$ it is defined by

$$\operatorname{Kurt} Z = \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^4]}{(\operatorname{Var} Z)^2}.$$

Note that $\operatorname{Kurt} Z \geq 1$ by Jensen's inequality. Further, kurtosis is invariant under affine transformations, that is

$$\operatorname{Kurt}(aZ + b) = \operatorname{Kurt}(Z).$$

This property is of great importance in regards to machine learning, because this means that the typical pre-processing steps of centering and dividing by the standard deviation do not affect the kurtosis of the features (or labels). In other words, the presence of $\operatorname{Kurt} \boldsymbol{x}$ in the expression for $\Sigma(\theta)$ cannot be explained away by a standardization of $\boldsymbol{x}$.

For convenience, here is a list of common distributions and their kurtosises.

Dist.	$\operatorname{Exp}(\lambda)$	$\operatorname{Poi}(\lambda)$	χ^2_n	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{U}[a, b]$	$\operatorname{Lognormal}(\mu, \sigma^2)$
Kurt.	9	$3 + \frac{1}{\lambda}$	$3 + \frac{12}{n}$	3	$\frac{9}{5}$	$e^{4\sigma^2} + 3e^{3\sigma^2} + 3e^{2\sigma^2} - 3$

Further, if $p \in [0, 1]$ and $Z \sim \operatorname{Bin}(1, p)$, then

$$\operatorname{Kurt} Z = \frac{3p^2 - 3p + 1}{p(1 - p)}$$

which has minimum 1 at $\frac{1}{2}$. That is, a symmetric Bernoulli attains the smallest possible Kurtosis of 1.

If $\operatorname{Kurt} Z = 3$, then we say Z (or its distribution) is *mesokurtic*. If $\operatorname{Kurt} Z > 3$, then Z is called *platykurtic* and we call Z *leptokurtic* for $\operatorname{Kurt} Z < 3$. These terms also delineate the settings for the error expansions in Section 3.2.1.

Finally, we remark that the common interpretation of kurtosis as heaviness of the tails of a distribution is somewhat misleading. Let us suppose the distribution of Z is unimodal, for simplicity. Then, according to Balanda and MacGillivray (1988), kurtosis is "vaguely [...] the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails [...]", that is higher kurtosis implies *both* higher peakedness as well as heavier tails. The term *shoulders* refers roughly to the area between the tails and the center. For multimodal distributions, the interpretation of kurtosis is a lot more involved or perhaps not even well understood. We will restrict our attention to unimodal distributions only (which includes all previous examples).

Appendix B. Results from (Stochastic) Analysis

Here we collect some known results from stochastic analysis that are needed for the proofs of our main theorems. We adapt the presentation to our setting in order to make the present article more self-contained.

Theorem 26 *Let $b, \sigma \in G_1(\mathbb{R}^d) \cap \mathrm{Lip}$, such that b is $\mathbb{R}^d$-valued and σ is $\mathbb{R}^{d\times d}$-valued. Then, for every $p \geq 2, T > 0$ and random field $\varphi : \Omega \times [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ with $\|\varphi^*\|_{p,T} < \infty$, the stochastic differential equation*

$$dX_t = b_t(X_t)\, dt + \sigma_t(X_t)\, dW_t, \quad X_0 = \varphi$$

admits a unique[8] *solution X on $[0,T]$, such that the family of solutions $X = (X_t)_{t\geq 0}$ satisfies $\|X^*\|_{p,T} < \infty$ and*

$$\|X^*\|_{p,T} \leq (1 + \|\varphi^*\|_{p,T}).$$

The same bound holds if we consider I-indexed families b, σ, φ and X for some index set I.

Proof This essentially a standard result (see, for example, Kunita, 2004, Theorems 3.1 and 3.2). The extension to an index set I and from an initial value $x \in \mathbb{R}^d$ to a process φ is discussed by Li et al. (2019, Theorem 18 and 19). ■

A (unordered) *multi-index* $\alpha \subseteq \{1, \dots, d\}$ is a multi-subset of $\{1, \dots, d\}$, that is a function $\alpha : \{1, \dots, d\} \to \mathbb{N}_0$. The size $\#\alpha$ of α is given by

$$\#\alpha := \sum_{j=1}^{d} \alpha(j).$$

Every subset $A \subseteq \{1, \dots, d\}$ becomes a multi-set by identifying it with its indicator function. Given multi-indices α and β we write $\alpha \leq \beta$ if $\alpha(j) \leq \beta(j)$ for all $j \in \{1, \dots, d\}$ and in that case the multi-index $\beta - \alpha$ is well defined by component-wise. Further, we write $j \in \alpha$ if $\{j\} \leq \alpha$ and set $\alpha - j := \alpha - \{j\}$ in that case.

If a function $f : \mathbb{R}^d \to \mathbb{R}$ is l-times continuously differentiable, then by Schwarz's theorem the partial derivative with respect to a multi-index α with $\#\alpha \leq l$ is well-defined recursively by

$$\partial_\alpha f = \partial_j \partial_{\alpha - j} f, \partial_\varnothing f = f.$$

where j is any $j \in \{1, \dots, d\}$ with $j \in \alpha$. Given $x \in \mathbb{R}^d$ and a multi-index α we define

$$x^\alpha := \prod_{j=1}^{d} x_j^{\alpha(j)}.$$

Theorem 27 *Let $l \in \mathbb{N}, p \geq 1$ and $b, \sigma \in G_1(\mathbb{R}^d) \cap \mathrm{Lip}^l$, such that b is $\mathbb{R}^d$-valued and σ is $\mathbb{R}^{d\times d}$-valued. Let $x \in \mathbb{R}^d, s \in [0,T]$ and X be the unique solution to the family of stochastic differential equations*

$$dX_t = b_t(X_t)\, dt + \sigma_t(X_t)\, dW_t, \quad X_s = x.$$

8. Of course, here we imply uniqueness up to indistinguishability.

Then X is l-times continuously differentiable w.r.t. x at any $(t,x) \in [s,T] \times \mathbb{R}^d$, a.s. and for every multi-index α with $0 < \#\alpha \leq l$, $\partial_\alpha X$ satisfies the stochastic differential equation

$$\partial_\alpha X_t = \psi_\alpha + \int_s^t \nabla b_u(X_u)\partial_\alpha X_u \, du + \int_s^t \nabla \sigma_u(X_u)\partial_\alpha X_u \, dW_u,$$

where $\|\psi_\alpha^*\|_{p,T} \in G(\mathbb{R}^d)$ for all $p \geq 2$. Moreover,

$$\mathbb{E}[\partial_\alpha X_t] = \partial_\alpha \mathbb{E}[X_t],$$

for all $t \geq 0$. Again, the results extend readily to I-indexed coefficients and processes for some index set I.

Proof For the proof we refer to Kunita (2004, Theorem 3.4). More specifically, for every $l \in \mathbb{N}$, assuming the result holds for all $l' < l$ define

$$Y := (X, \partial_1 X, \dots \partial_d X, \partial_{1,1} X, \dots, \partial_{1,d} X, \partial_{2,1} X, \dots, \partial_{d,\dots,d} X)^\dagger,$$

where the last partial derivative is of the order $l-1$. Then Y satisfies the stochastic differential equation

$$\begin{aligned} Y = \begin{pmatrix} x \\ e_1 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \psi_1 \\ \vdots \\ \psi_{d,\dots,d} \end{pmatrix} + \int_s^t \begin{pmatrix} b_u(X_u) \\ \nabla b_u(X_u)\partial_1 X_u \\ \vdots \\ \nabla^{l-1} b_u(X_u)\partial_{d,\dots,d} X_u \end{pmatrix} du \\ + \int_s^t \begin{pmatrix} \sigma_u(X_u) \\ \nabla \sigma_u(X_u)\partial_1 X_u \\ \vdots \\ \nabla^{l-1} \sigma_u(X_u)\partial_{d,\dots,d} X_u \end{pmatrix} dW_u, \end{aligned}$$

where the processes $\psi_1, \dots, \psi_{d,\dots,d}$ consist of additional integrals $\int_s^t du$ and $\int_s^t dW_u$ of the remaining terms induced by repeated application of the chain rule. The terms within $\int_s^t du$ and $\int_s^t dW_u$ respectively are seen to be functions of u and the state Y, satisfying the conditions given by Kunita (2004, Theorem 3.4). By applying it again to the SDE governing Y the result follows via induction on l. ■

We denote the spectral norm of a square matrix or linear map A by $\|A\|_2$. Note that for any linear map $\Phi : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$, where $\mathbb{R}^{d\times k}$ is equipped with the Frobenius norm, the operator norm is given by

$$\|\Phi\|_{\text{op}}^2 = \sup_{\|A\|_F=1} \|\Phi A\|_F^2 = \sup_{\|A\|_F=1} \langle \Phi^\dagger \Phi K, K\rangle = \lambda_{\max}(\Phi^\dagger \Phi) = \|\Phi\|_2^2,$$

where $\Phi^\dagger$ is the adjoint operator of Φ. Hence,

$$\|\Phi K\|_F \leq \|\Phi\|_2 \|K\|_F, \quad K \in \mathbb{R}^{d\times k}.$$

Lemma 28 *Let $A : [0,T] \to \mathbb{R}^{d\times d}$ be a family of symmetric matrices and $\Phi : [0,T] \to \mathbb{R}^{d\times d} \in C^1$ a solution to the non-autonomous matrix linear ODE*

$$d\Phi_t = A_t\Phi_t\, dt, \quad \Phi_0 = 1_{d\times d}.$$

Then,

$$\|\Phi_t\Phi_s^{-1}\|_2 \leq \exp\left(\int_s^t \lambda_{\max}(A_r)\, dr\right), \quad 0 \leq s \leq t \leq T.$$

This is a combination of special cases of Lemma 1b and 1c by Strom (1975).

Theorem 29 *Let $b : \mathbb{R}^d \to \mathbb{R}^d \in \mathrm{Lip}^2$ and $X : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$ be the unique solution to the ordinary differential equation*

$$dX_t = b(X_t)\, dt,$$

in the sense that

$$dX_t(x) = b(X_t(x))\, dt, \quad X_0(x) = x,$$

for all $x \in \mathbb{R}^d$. Then $X : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d \in C^2([0,T]\times\mathbb{R}^d)$, with

$$d\nabla X_t = \nabla b(X_t)\nabla X_t\, dt, \quad \nabla X_0 = 1_{d\times d}, \tag{43}$$

and

$$d\nabla^2 X_t^k = (\nabla X_t)^\dagger \nabla^2 b_k(X_t)\nabla X_t + \sum_{l=1}^d \partial_l b_k(X_t)\nabla^2 X_t^l\, dt, \quad \nabla^2 X_0^k = 0, \quad k \in \{1,\dots,d\}. \tag{44}$$

Further, suppose $b = -\nabla\mathcal{R}$ and define

$$M_t := \exp\left(-\int_0^t \lambda_{\min}(\nabla^2\mathcal{R}(X_s))\, ds\right).$$

Then, $\|\nabla X_t\|_2 \leq M_t$, and

$$\|\nabla^2 X_t\|_F \leq dM_t \int_0^t M_s \|\nabla^3\mathcal{R}(X_s)\|_F\, ds, \tag{45}$$

for all $t \in [0,T]$.

Proof Since $b \in \mathrm{Lip}^2$, we have $X : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d \in C^2([0,T]\times\mathbb{R}^d)$ (see Teschl, 2012, Theorem 2.10). Since higher partial derivatives can be exchanged, we have

$$\partial_t\partial_j X_t^i = \partial_j\partial_t X_t^i = \sum_{k=1}^d \partial_k b_i(X_t)\partial_j X_t^k, \quad i,j \in \{1,\dots,d\},$$

and hence (43) follows. Now, let us consider $\nabla^2 X$. We compute

$$\partial_t\partial_{ij} X_t^k = \partial_i\left(\sum_{l=1}^d \partial_l b_k(X_t)\partial_j X_t^l\right) = \sum_{m,l}^d \partial_{ml} b_k(X_t)\partial_i X_t^m \partial_j X_t^l + \sum_{l=1}^d \partial_l b_k(X_t)\partial_{ij} X_t^l,$$

for $i, j, k \in \{1, \dots, d\}$. Thus, (44) holds true. Now, assume $b = -\nabla \mathcal{R}$. Lemma 28 implies $\|\nabla X_t\|_2 \leq M_t$. Denote by $\mathcal{L}(V, V)$ the set of linear operators $V \to V$ for some vector space V. Define a family of linear operators $A : [0, T] \to \mathcal{L}(\mathbb{R}^{d\times d\times d}, \mathbb{R}^{d\times d\times d})$ by

$$A_t(K)_{ijk} = \sum_{l=1}^{d} \partial_l \partial_i \mathcal{R}(X_t) K_{ljk}, \quad t \in [0, T], K \in \mathbb{R}^{d\times d\times d}, i, j, k \in \{1, \dots, d\}.$$

Further, define $f : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d\times d}$ by

$$(f_t)_{ijk} = ((\nabla X_t)^\dagger \nabla^3 \mathcal{R}_{\cdot,\cdot,k}(X_t) \nabla X_t)_{ij}, \quad t \in [0, T], i, j, k \in \{1, \dots, d\}.$$

Recall that $\nabla^2 X_{ijk} = \partial_{jk} X^i$. Therefore, we can write (44) as

$$d\nabla^2 X_t = -A_t(\nabla^2 X_t) - f_t. \tag{46}$$

Fix $t \geq 0$. Let $\{v^1, \dots, v^d\}$ be an eigenbasis of $\nabla^2 \mathcal{R}(X_t)$ with corresponding eigenvalues $\lambda_1, \dots, \lambda_d$. Then for all $i, j, k \in \{1, \dots, d\}$, define $K^{i,j,k} \in \mathbb{R}^{d\times d\times d}$ by

$$K^{i,j,k}_{l,m,n} = v^i_l \delta_{j,m} \delta_{k,n},$$

where δ is the Kronecker delta. Then (see also user1551 on Math Stack Exchange, 2023),

$$\begin{aligned} A_t(K^{ijk})_{lmn} =& \sum_{p=1}^{d} \nabla^2 \mathcal{R}(X_t)_{lp} K^{ijk}_{pmn} = \sum_{p=1}^{d} \nabla^2 \mathcal{R}(X_t)_{lp} K^{ijk}_{pmn} \\ =& \lambda_i v^i_l \delta_{j,m} \delta_{k,n} = \lambda_i K^{ijk}_{lmn}. \end{aligned}$$

Thus, $K^{i,j,k}$ is an eigenvector of A_t with corresponding eigenvalue λ_i. Further, the $K^{i,j,k}$ are linearly independent and thus form an eigenbasis of A_t. In particular, the set of eigenvalues for A_t and $\nabla^2 \mathcal{R}(X_t)$ coincide, and we have $\lambda_{\min}(A_t) = \lambda_{\min}(\nabla^2 \mathcal{R}(X_t))$. Further, suppose $\Phi : [0, T] \to \mathcal{L}(\mathbb{R}^{d\times d\times d}, \mathbb{R}^{d\times d\times d})$ is a solution to the operator-valued ODE

$$d\Phi_t = -A_t(\Phi_t), \quad \Phi_0 = \mathrm{id}_{\mathbb{R}^{d\times d\times d}}.$$

Then, by Lemma 28,

$$\|\Phi_t \Phi_s^{-1}\|_2 = \|\tilde{\Phi}_t \tilde{\Phi}_s^{-1}\|_2 \leq \exp\left(\int_s^t \lambda_{\max}(-\nabla^2 \mathcal{R}(X_s))\, ds\right) = \frac{M_t}{M_s},$$

where $\tilde{\Phi}_t$ is a matrix representing Φ_t in some basis of $\mathbb{R}^{d\times d\times d}$. Further, the solution to (46) is given by

$$\nabla^2 X_t = \int_0^t \Phi_t \Phi_s^{-1} f_s\, ds.$$

We have

$$\|f_t\|_F^2 = \sum_{k=1}^{d} \|(f_t)_{\cdot\cdot k}\|_F^2 \leq \|\nabla X_t\|_F^4 \sum_{k=1}^{d} \|\nabla^3 \mathcal{R}_{\cdot,\cdot,k}(X_t)\|_F^2 = \|\nabla X_t\|_F^4 \|\nabla^3 \mathcal{R}(X_t)\|_F^2,$$

and so

$$\begin{aligned}\|\nabla^2 X_t\|_F \leq& \int_0^t \|\Phi_t \Phi_s^{-1}\|_2 \|f_s\|_F \, ds \\ \leq& \int_0^t \frac{M_t}{M_s} \|\nabla X_s\|_F^2 \|\nabla^3 \mathcal{R}(X_s)\|_F \, ds \\ \leq& d M_t \int_0^t M_s \|\nabla^3 \mathcal{R}(X_s)\|_F \, ds,\end{aligned}$$

noting that $\|A\|_F \leq \sqrt{d}\|A\|_2$ for any square matrix A. ■

Given a set A the *Kleene closure* is the set of all A-tuples of arbitrary length, that is

$$A^* := \bigcup_{n \geq 0} A^n,$$

where $A^0 = \{()\}$. We let $|(a_1, \ldots, a_n)| = n$ and $|()| = 0$ be the *length* of such a tuple.

We care about the set of (ordered) *multi-indices* $\{0, \ldots, d\}^*$, where $\mathbb{R}^d$ is the state space of W. Note that we have $(1, 2) \neq (2, 1)$, unlike the case of (unordered) multi-indices considered before. Given a multi-index $\alpha \in \{0, \ldots, d\}^*$ of length $l = |\alpha| > 0$ we define the *left-* and *right deletions*

$$\alpha^- = (\alpha_1, \ldots, \alpha_{l-1}), {}^-\alpha = (\alpha_2, \ldots, \alpha_l) \in \{0, \ldots, d\}^{l-1}.$$

Let $\mathcal{H}^{()}$ be the set of all continuous stochastic processes and define

$$\begin{aligned}\mathcal{H}^{(0)} &= \{X \in \mathcal{H}^{()} : \int_0^t |X_s| \, ds < \infty, a.s., t \geq 0\}, \\ \mathcal{H}^{(1)} &= \{X \in \mathcal{H}^{()} : \int_0^t |X_s|^2 \, ds < \infty, a.s., t \geq 0\}.\end{aligned}$$

Also for convenience set $\mathcal{H}^{(j)} := \mathcal{H}^{(1)}$ for all $j \in \{1, \ldots, d\}$.

We let $W_t^0 = t, t \geq 0$. Given a progressively measurable stochastic process $X : \Omega \times [0, \infty) \to \mathbb{R}^d$ and $\alpha \in \{0, \ldots d\}^*$ with $l = |\alpha|$ we define the *multiple Itô integral*

$$\int_s^t X \, dW^\alpha = \begin{cases} X, & |\alpha| = 0, \\ \int_s^t \int_s^u X \, dW^{\alpha^-} \, dW^{\alpha_l}, & |\alpha| > 0, \end{cases}$$

as long as $X \in \mathcal{H}^\alpha$, where the latter is the case exactly when

$$\int_s^\cdot X \, dW^{\alpha-} = \left(\int_s^t X \, dW^{\alpha-} \right)_{t \geq s} \in \mathcal{H}^{(\alpha_l)}.$$

Further, given $f \in C^{1,2}([0,\infty) \times \mathbb{R}^d)$ define

$$\mathcal{A}_X f := \mathcal{L}^0 f := \frac{\partial f}{\partial t} + \nabla f^\dagger b + \frac{1}{2}\operatorname{tr}(\nabla^2 f \sigma\sigma^\dagger),$$
$$\mathcal{L}^j f := \sigma_{j,\cdot}^\dagger \nabla f = \sum_{k=1}^d \sigma_{k,j} \partial_{x_k} f, j \in \{1,\dots,d\}.$$

For any $\alpha \in \{0,\dots,d\}^*$ set

$$\alpha(0) := \#\{j : \alpha_j = 0\}.$$

Given $f \in C^{\alpha(0),2(|\alpha|-\alpha(0))}([0,\infty) \times \mathbb{R}^d)$ we define the *Itô coefficient function*

$$\mathcal{L}^\alpha f := \begin{cases} f, & |\alpha| = 0, \\ \mathcal{L}^{\alpha_1}(\mathcal{L}^{-\alpha} f), & |\alpha| > 0. \end{cases}$$

Theorem 30 *Let $b, \sigma \in G_1(\mathbb{R}^d) \cap \mathrm{Lip}$, such that b is $\mathbb{R}^d$-valued and σ is $\mathbb{R}^{d\times d}$-valued, $0 \le s \le t \le T, x \in \mathbb{R}^d$ and let X be the unique solution to the stochastic differential equation*

$$dX_t = b_t(X_t)\,dt + \sigma_t(X_t)\,dW_t, \quad X_s = x.$$

on $[s,T]$. Then given $f \in C^{\alpha(0),2(|\alpha|-\alpha(0))}([0,\infty) \times \mathbb{R}^d)$ we have

$$f(T, X_T) = \sum_{|\alpha| \le l} \int_s^T \mathcal{L}^\alpha f(s, X_s)\,dW^\alpha + \sum_{|\beta| = l+1} \int_s^T \mathcal{L}^\alpha f(\cdot, X_\cdot)\,dW^\alpha.$$

Further, applying expectation yields

$$\mathbb{E} f(T, X_T) = \sum_{i=0}^{l} \frac{(T-s)^i}{i!} \mathcal{A}_X^i f(s, X_s) + \int_s^T \int_s^{u_1} \cdots \int_s^{u_l} \mathbb{E}\mathcal{A}_X^{l+1} f(u_{l+1}, X_{u_{l+1}})\,du_{l+1} \dots du_1.$$

Proof We refer to Kloeden and Platen (1995, Theorem 5.5.1. page 182). All the iterated integrals are defined since $\mathcal{L}^\alpha f(\cdot, X_\cdot) \in \mathcal{H}^\alpha$ for all α with $|\alpha| \le l$. As the hierarchical set choose $\mathcal{A} := \{\alpha : |\alpha| \le l\}$. For the second statement note that

$$\int_s^T \int_s^{u_1} \cdots \int_s^{u_{i-1}} 1\,du_i \dots du_1 = \frac{1}{i!}(T-s)^i,$$

and that any integral $\int_s^T dW^\alpha$ with $\alpha(0) < |\alpha|$ has expectation zero. ■

Lemma 31 *Consider the stochastic differential equation*

$$dX_t = b_t(X_t)\,dt + \sigma_t(X_t)\,dW_t,$$

where

$$b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d, \sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d \times d} \in G_1([0,T] \times \mathbb{R}^d) \cap \mathrm{Lip}$$

and additionally

$$b, \sigma \in G^l([0,T] \times \mathbb{R}^d) \cap C^{l',l}([0,T] \times \mathbb{R}^d).$$

Let $f : [0,T] \times \mathbb{R}^d \to \mathbb{R} \in G^l([0,T] \times \mathbb{R}^d) \cap C^{l',l}([0,T] \times \mathbb{R}^d)$. *Then,*

$$\mathcal{A}_X^i f \in G^{l-2i} \cap C^{l'-i,l-2i}([0,T] \times \mathbb{R}^d),$$

for all $i \in \mathbb{N}$ *with* $i \leq \frac{l}{2} \wedge l'$, *where* $\mathcal{A}_X$ *is the infinitesimal generator of* X.

Proof Suppose the statement holds for all $i' < i$. Then $\mathcal{A}_X^i f = \mathcal{A}_X g$ for some

$$g \in C^{l'-(i-1),l-2(i-1)}([0,T] \times \mathbb{R}^d)$$

with $g \in G^{l-2(i-1)}(\mathbb{R}^d)$. Then,

$$b^\dagger \nabla g \in G^{l-2i+1}(\mathbb{R}^d), \quad \mathrm{tr}[\sigma^\dagger \sigma \nabla^2 g] = \sum_{j,k}^{d} (\sigma^\dagger \sigma)_{j,k} \partial_{j,k} g \in G^{l-2i}(\mathbb{R}^d),$$

and $\partial_t g \in C^{l'-i,l-2i+2}([0,T] \times \mathbb{R}^d)$. Combining all three statements yields the result. ■

References

A. Ali, E. Dobriban, and R. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine learning*, pages 233–244. PMLR, 2020.

J. An, J. Lu, and L. Ying. Stochastic modified equations for the asynchronous stochastic gradient descent. *Information and Inference: A Journal of the IMA*, 9(4):851–873, 2019.

K. P. Balanda and H. L. MacGillivray. Kurtosis: A critical review. *The American Statistician*, 42(2):111–119, 1988. URL `http://www.jstor.org/stable/2684482`.

N. M. Boffi and J.-J. E. Slotine. A continuous-time analysis of distributed stochastic gradient. *Neural computation*, 32(1):36–96, 2020.

A. Bose. *Random Matrices and Non-Commutative Probability*. CRC Press LLC, 2021. ISBN 9780367700812.

H. Cao and X. Guo. Approximation and convergence of GANs training: An SDE approach. *arXiv preprint arXiv:2006.02047*, 2020.

P. Chen, Q.-M. Shao, and L. Xu. A universal probability approximation method: Markov process approach. *arXiv preprint arXiv:2011.10985*, 2020.

O. Elkabetz and N. Cohen. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34:4947–4960, 2021.

Y. Feng, L. Li, and J.-G. Liu. Semi-groups of stochastic gradient descent and online principal component analysis: Properties and diffusion approximations. *arXiv preprint arXiv:1712.06509*, 2018.

Y. Feng, T. Gao, L. Li, J.-G. Liu, and Y. Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *arXiv preprint arXiv:1902.00635*, 2019.

D. A. Fetisov. On symmetric solutions to linear matrix time-varying differential equations. *Journal of Physics: Conference Series*, 2090(1):012134, 2021. Publisher: IOP Publishing.

X. Fontaine, V. De Bortoli, and A. Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

C. Graham and D. Talay. *Stochastic Simulation and Monte Carlo methods: Mathematical Foundations of Stochastic Simulation*, volume 68. Springer Science & Business Media, 2013.

H. Gu, X. Guo, and X. Li. Adversarial training for gradient descent: Analysis through its continuous-time approximation. *arXiv preprint arXiv:2105.08037*, 2023.

G. Hu and Y. Zhang. Runtime analysis of stochastic gradient descent. In A. Emrouznejad and J. R. Chou, editors, *CSAE 2020: The 4th International Conference on Computer Science and Application Engineering, Sanya, China, October 20-22, 2020*, pages 15:1–15:6. ACM, 2020.

W. Hu, C. J. Li, L. Li, and J.-G. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 1995.

D. Kunin, J. Sagastuy-Brena, L. Gillespie, E. Margalit, H. Tanaka, S. Ganguli, and D. L. K. Yamins. Rethinking the limiting dynamics of SGD: modified loss, phase space oscillations, and anomalous diffusion. 2022. URL `https://openreview.net/forum?id=mRc_t2b3l1-`.

H. Kunita. Stochastic differential equations based on levy processes and stochastic flows of diffeomorphisms. In *Real and Stochastic Analysis : New Perspectives*. Birkhäuser Boston, Boston, MA, 2004. ISBN 1461220548.

A. Lanconelli and C. S. A. Lauria. A note on diffusion limits for stochastic gradient descent. *arXiv preprint arXiv:2210.11257*, 2022.

Q. Li, C. Tai, and E. Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

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

Z. Li, K. Lyu, and S. Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. In *Advances in Neural Information Processing Systems*, volume 33, pages 14544–14555. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a7453a5f026fb6831d68bdc9cb0edcae-Abstract.html.

Z. Li, S. Malladi, and S. Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). *Advances in Neural Information Processing Systems*, 34:12712–12725, 2021.

Z. Li, T. Wang, and S. Arora. What happens after SGD reaches zero loss? – A mathematical framework. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=siCt4xZn5Ve.

S. Malladi, K. Lyu, A. Panigrahi, and S. Arora. On the SDEs and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35:7697–7711, 2022.

S. Mandt, M. D. Hoffman, and D. M. Blei. Continuous-time limit of stochastic gradient descent revisited. *8th International Workshop on "Optimization for Machine Learning"*, 2015.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f4661398cb1a3abd3ffe58600bf11322-Abstract.html.

T. Strom. On logarithmic norms. *SIAM Journal on Numerical Analysis*, 12(5):741–753, 1975. URL https://www.jstor.org/stable/2156188. Publisher: Society for Industrial and Applied Mathematics.

D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.

G. Teschl. *Ordinary Differential Equations and Dynamical Systems*. Graduate studies in mathematics. American Mathematical Society, 2012. ISBN 9780821883280.

user1551 on Math Stack Exchange. Eigenvalues of linear map of 3-tensors given by matrix, 2023. URL https://math.stackexchange.com/q/4747936. URL:https://math.stackexchange.com/q/4747936 (version: 2023-08-05).

S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type part ii: Continuous time analysis. *Journal of Nonlinear Science*, 34(1):1–45, 2024.

Z. Xie, I. Sato, and M. Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

J. Yang, W. Hu, and C. J. Li. On the fast convergence of random perturbations of the gradient flow. *Asymptotic Analysis*, 122(3-4):371–393, 2021.

P. Zhou, J. Feng, C. Ma, C. Xiong, S. C. H. Hoi, et al. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.