A comparison of continuous-time approximations to stochastic gradient descent

Stefan Ankirchner, Stefan Perko

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

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 leading 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. Further, we demonstrate that for Gaussian features both SDE approximations are equally good. However, for leptokurtic features we find that the SDE approximation with state-dependent diffusion coefficient is of higher quality than the approximation with state-independent noise. Moreover, the relationship reverses for platykurtic features.

Keywords: Stochastic gradient descent, gradient flow, stochastic differential equation, weak approximation, learning rate schedules, Talay-Tubaro expansion

1 Introduction

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

$$\chi^h_{n+1} = \chi^h_n - h\nabla R_{\gamma(n)}(\chi^h_n), \quad n \in \mathbb{N}_0, \tag{1}$$

where $(R_r)_{r\in\Gamma}$ is a family of differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$, h is a positive real, and $(\gamma(n))_{n\in\mathbb{N}_0}$ is an i.i.d. sequence of Γ-valued 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 sample of the data 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

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

as $h \downarrow 0$.

GF dynamics are deterministic and hence ignore the randomness in SGD. Therefore, in recent years analytic approximations in terms stochastic differential equations (SDEs) have become common. In particular, SDE approximations have been used to optimize hyperparameters (see, for example, Mandt et al. (2015), Mandt et al. (2017), Li et al. (2017), Malladi et al. (2022)), to analyze the long-term behavior of SGD processes (see, for example, Cao and Guo (2020), Kunin et al. (2022), Wojtowytsch (2021)), to study the impact of normalization schemes (see, for example, Li et al. (2020)), to analyze the runtime until convergence (see, for example, Hu and Zhang (2020)), to study the transition between stationary points (see, for example, Yang et al. (2020), Zhou et al. (2020), Xie et al. (2020), Hu et al. (2017)), to study the implicit bias and regularization properties of SGD (Ali et al. (2020), Pesme et al. (2021), Li et al. (2022)) and to study the effect of running SGD in parallel (see, for example, 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 positive semi-definite matrix. Two choices for σ are common: first, σ is constant, that is independent of the state (see for example 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 for example 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 leading error terms, that is the constants in front of the linear term in the error expansion. It turns out that the leading error terms for GF, CC-SGF and NCC-SGF are all different. We can thus confirm a conjecture proposed in Feng et al. (2018) (Remark 2.3.).

We characterize the leading 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 in Talay and Tubaro (1990) refers to a *single* SDE.

We show that for linear regression models, the leading error terms can be calculated in closed form. A comparison then reveals that one can always reduce the leading term of the GF approximation by introducing a diffusion term. Moreover, there is not a clear favorite among the SDE approximations: the preferred approximation depends, surprisingly, on the kurtosis of the features. Now, we provide a more detailed summary of our contributions.

Summary of contributions

- 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 leading error term in the error expansion with respect to the learning rate.

- Using the leading error term expansion we show that in the example of linear regression with Gaussian features and non-zero residuals, that is data noise, using population risk as test function, the GF approximation is always inferior to both the SGF approximations. Moreover, the SGF approximations have the same approximation quality, which justifies usage of the simpler constant covariance SGF.

- Finally, in the non-Gaussian setting we show that the NCC-SGF is best, among the three approximations, for leptokurtic features, while the CC-SGF approximation is superior for platykurtic features. Moreover, in the case of kurtosis 2 the quality of the CC-SGF approximation jumps from order 1 to order 2, for the specific population risk test function.

Related Work The idea to use stochastic differential equations for approximating SGD processes appears first in Mandt et al. (2015), Li et al. (2017) and Li et al. (2019). In Mandt et al. (2015) the authors heuristically use CC-SGF for approximating and analyzing the SGD process. Li et al. (2017) derive NCC-SGF and rigorously prove in Li et al. (2019) that it is a first-order approximation of SGD. 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 leading error term explicitly.

Further results for the NCC-SGF approximation include Lanconelli and Lauria (2022), Chen et al. (2020) and Fontaine et al. (2021). In Lanconelli and Lauria (2022) the NCC-SGF dynamics are justified with a general Markov chain convergence theorem. Theorem 3.5. in Chen et al. (2020) provides an estimate of the Wasserstein-1 distance between SGD processes and NCC-SGF. The article 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 $\alpha \in [0, 1)$. (Fontaine et al., 2021, Proposition 25)

provides an asymptotic estimate of the weak error as γ converges to zero. 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 leading error terms and we do not make a specific assumption on the learning rate schedule u.

Moreover, the literature comprises articles considering weak approximations of order 2 for SGD processess (for example Li et al. (2019), Feng et al. (2018), Feng et al. (2019) and Gu and Guo (2021)).

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 in Li et al. (2021).

2 General results on leading 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 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 sample 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 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 $h \in (0, 1)$ we consider the sequence of learning rates $(hu_{nh})_{n\in\mathbb{N}_0}$. The parameter $h \in (0, 1)$ acts as discretization parameter and can be interpreted as the *maximal* learning rate if u is not constant.

Recall that γ maps into 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

$$\chi^h_{n+1} = \chi^h_n + hu_{nh}H_{\gamma(n)}(\chi^h_n), \quad \chi_0 = x. \tag{5}$$

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

The prototypical example to keep in mind is online SGD with replacament. 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 positive semi-definite and symmetric, 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.*

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.

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 $X_t^t(x) = x$. 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

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

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.

1. See the appendix before Theorem 22 for a definition of (unordered) multi-indices.
2. Here, ∇ denotes the gradient and ∇^2 the Hessian matrix with respect to x.

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

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),

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

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 in 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 $h \in \mathcal{H}$,*

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

where φ is defined in (10).

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

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. Hence, let $\theta^* \in \mathbb{R}^d$ with $\bar{H}(\theta^*) = 0$. Then for every $h \in \mathcal{H} \cup \{0\}$ we consider the SDE

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

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 in Ali et al. (2020)), with the small caveat that each stationary point θ^* yields a different CC-SGF. In the case $u = 1$ this is essentially the continuous-time approximation introduced by Mandt et al. (2015).

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

Theorem 3 *Assume (A1), (A2) and (A3). 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 $h \in \mathcal{H}$,*

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

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

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.

Firstly, we provide a theoretical comparison using Theorems 1, 2 and 3. In the multi-dimensional setting with Gaussian features we observe that GF is always a poorer approximation than SGF in the presence of noisy data, that is non-zero residuals. Somewhat surprisingly, the quality of approximation for NCC-SGF and CC-SGF turn out to be the same. Afterwards we take a look at non-Gaussian features in dimension 1. We observe that in this case NCC-SGF is a better approximation to SGD for *leptokurtic* features. In contrast, for *platykurtic* features CC-SGF is better.

Secondly, we substantiate these 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 of i.i.d. data. Note, however, that by choosing ν to be an empirical measure, we recover the setting of SGD with replacement for a finite set of data. In this case ν represents a *sample* rather than the population and $\mathcal{R}$ should instead be interpreted as *empirical risk*, more commonly known as the *training error*.

Notation In the remainder of the section we use the following notation. 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

$$\langle u, v\rangle = u^\dagger v, \quad \langle u, v\rangle^2 = \langle u^{\otimes 2}, v^{\otimes 2}\rangle \text{ and } \langle A, B\rangle = \operatorname{tr}(A^\dagger B) \in \mathbb{R}$$

The quantity $\langle A, B\rangle$ is also known as the Frobenius inner product of A and B. Note that for matrices $A, B, C \in \mathbb{R}^{d\times d}$, we have

$$\langle A, BC\rangle = \operatorname{tr}(A^\dagger BC) = \langle B^\dagger A, C\rangle = \operatorname{tr}(CA^\dagger B) = \langle AC^\dagger, B\rangle.$$

Further, 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 $B^{\otimes 2} := B \otimes B$. In particular, $u^{\otimes 2} = uu^\dagger$ for any $u \in \mathbb{R}^d$.

3.1 The statistical learning setting

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, $\mathbb{E}\boldsymbol{\varepsilon} = 0, \sigma_\varepsilon^2 := \mathbb{E}\boldsymbol{\varepsilon}^2 < \infty$ 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.

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

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

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^*. We can determine an estimate of θ^* using stochastic gradient descent

$$\chi_{n+1}^h = \chi_n^h - h \nabla_\theta R_{\boldsymbol{x}_n,\boldsymbol{y}_n}(\chi_n^h) = \chi_n^h - h(\langle \chi_n^h, \boldsymbol{x}_n \rangle - \boldsymbol{y}_n)\boldsymbol{x}_n, \tag{16}$$

where $(\boldsymbol{x}_n, \boldsymbol{y}_n)_{n \in \mathbb{N}_0}$ is an i.i.d. sequence with $(\boldsymbol{x}_0, \boldsymbol{y}_0) \sim \nu$. For simplicity we only consider constant learning rates in this section. We calculate

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

where $\kappa := \operatorname{Cov} \boldsymbol{x}$ is the covariance matrix of $\boldsymbol{x}$.

We define the covariance matrix of the gradient noise by

$$\Sigma(\theta) := \mathrm{Cov}[\nabla_\theta R_{\boldsymbol{x},\boldsymbol{y}}(\theta)].$$

Then,

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

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

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

3.2 Theoretical comparison for Gaussian features

In this section we compare GF with its stochastic counterparts by deriving the leading error terms for the population risk of GF, NCC-SGF and CC-SGF explicitly.

We assume that the features are centered Gaussian, that is $x \sim \mathcal{N}(0, \kappa)$.

Let τ be permutation of l elements and $B \in \mathbb{R}^{d^{\times l}}$ an l-tensor. Then we write $B_\tau \in \mathbb{R}^{d^{\times l}}$ for

$$(B_\tau)_{i_1,\ldots,i_l} = B_{\tau(i_1),\ldots\tau(i_l)}.$$

For example if B is matrix, then $B^\dagger = B_{(12)}$. Here we use the cycle notation for permutations. By Isserli's theorem (see for example 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

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

Therefore, we can simplify the variance of the gradient noise to

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

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

$$\begin{aligned}
dX_t^0 =& -\kappa(X_t^0 - \theta^*)\,dt \\
dX_t^h =& -\kappa(X_t^h - \theta^*)\,dt + \sqrt{h}\sqrt{2\kappa(\theta-\theta^*)^{\otimes 2}\kappa + \sigma_\varepsilon^2\kappa}\,dW_t \\
dX_t^{\mathrm{CC},h} =& -\kappa(X_t^{\mathrm{CC},h} - \theta^*)\,dt + \sqrt{h\sigma_\varepsilon^2\kappa}\,dW_t.
\end{aligned} \tag{17}$$

Note that the process with constant covariance dynamics (17) is an Ornstein-Uhlenbeck process. By applying the results from Section 2 to the population risk we obtain the following.

Proposition 4 *Suppose $\chi_0^h = X_0^h = X_0^{\mathrm{CC},h} = \theta \in \mathbb{R}^d$ for all $h \in (0,1)$. Then, we have*

$$\begin{aligned}
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^0) =& \frac{h}{2} T \langle \kappa^3 e^{-2T\kappa}, (\theta-\theta^*)^{\otimes 2} \rangle + \frac{h}{2}\sigma_\varepsilon^2 \langle \kappa^2, \int_0^T e^{-2(T-t)\kappa}\, dt \rangle \\
&+ \mathcal{O}(h^2) \qquad (18)\\
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^h) =& -\frac{h}{2} T \langle \kappa^3 e^{-2T\kappa}, (\theta-\theta^*)^{\otimes 2} \rangle + \mathcal{O}(h^2) \\
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^{\mathrm{CC},h}) =& \frac{h}{2} T \langle \kappa^3 e^{-2T\kappa}, (\theta-\theta^*)^{\otimes 2} \rangle + \mathcal{O}(h^2),
\end{aligned}$$

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

Remark 5 *Note that if κ is positive definite, then the term with the integral in* (18) *can be simplified as follows:*

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

Proof of Proposition 4 Set $v_t(\theta) := \mathcal{R}(X_T^{0,t}(\theta))$. Then by Theorem 1, 2 and 3 we have, for $T > 0$,

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

as $\mathcal{H} \ni h \downarrow 0$, where

$$\begin{aligned}
\varphi =& \frac{1}{2}\langle \nabla^2 v, (\nabla \mathcal{R})^{\otimes 2} \rangle - \langle \partial_t \nabla v, \nabla \mathcal{R} \rangle + \frac{1}{2}\partial_t^2 v, \\
\varphi^{\mathrm{GF}} =& \varphi + \frac{1}{2}\langle \nabla^2 v, \Sigma \rangle, \\
\varphi^{\mathrm{CC}} =& \varphi + \frac{1}{2}\langle \nabla^2 v, \Sigma - \Sigma(\theta^*) \rangle.
\end{aligned}$$

Starting gradient flow in θ at t, we have

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

Write $a_t = e^{-\kappa(T-t)}$. Note that the matrices $\kappa, e^{b\kappa}$ and $e^{c\kappa}$ commute with each other for all $b, c \in \mathbb{R}$. Then,

$$\begin{aligned} v_t(\theta) =& \frac{1}{2}\langle \kappa, (X_T^{0,t}(\theta) - \theta^*)^{\otimes 2}\rangle + \frac{\sigma_\varepsilon^2}{2} \\ =& \frac{1}{2}\langle \kappa, (e^{-(T-t)\kappa}(\theta - \theta^*))^{\otimes 2}\rangle + \frac{\sigma_\varepsilon^2}{2} \\ =& \frac{1}{2}\langle \kappa, e^{-(T-t)\kappa}(\theta - \theta^*)^{\otimes 2} e^{-(T-t)\kappa}\rangle + \frac{\sigma_\varepsilon^2}{2}, \end{aligned}$$

and hence

$$\begin{aligned} \nabla v_t(\theta) =& \kappa a_t^2(\theta - \theta^*) \\ \nabla^2 v_t(\theta) =& \kappa a_t^2 \\ \langle \nabla^2 v_t(\theta), (\nabla \mathcal{R}(\theta))^{\otimes 2}\rangle =& \langle \kappa^3 a_t^2, (\theta - \theta^*)^{\otimes 2}\rangle \\ \langle \nabla^2 v_t(\theta), \Sigma(\theta) - \Sigma(\theta^*)\rangle =& 2\langle \kappa^3 a_t^2, (\theta - \theta^*)^{\otimes 2}\rangle \\ \langle \nabla^2 v_t(\theta), \Sigma(\theta)\rangle =& 2\langle \kappa^3 a_t^2, (\theta - \theta^*)^{\otimes 2}\rangle + \sigma_\varepsilon^2\langle \kappa^2, a_t^2\rangle \\ \partial_t \nabla v_t(\theta) =& 2\kappa^2 a_t(\theta - \theta^*) \\ -\langle \partial_t \nabla v_t(\theta), \nabla \mathcal{R}(\theta)\rangle =& -2\langle \kappa^3 a_t^2, (\theta - \theta^*)^{\otimes 2}\rangle \\ \partial_t v_t(\theta) =& \langle \kappa^2 a_t^2, (\theta - \theta^*)^{\otimes 2}\rangle \\ \partial_t^2 v_t(\theta) =& 2\langle \kappa^3 a_t^2, (\theta - \theta^*)^{\otimes 2}\rangle. \end{aligned}$$

Further,

$$(X_t^0(\theta) - \theta^*)^{\otimes 2} = e^{-t\kappa}(\theta - \theta^*)^{\otimes 2} e^{-t\kappa}$$

and $a_t e^{\kappa t} = e^{-\kappa T}$. Thus,

$$\begin{aligned} \varphi_t(X_t^0(\theta)) =& -\frac{1}{2}\langle \kappa^3 a_t^2, (X_t^0(\theta) - \theta^*)^{\otimes 2}\rangle \\ =& -\frac{1}{2}\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle \\ \varphi_t^{\mathrm{CC}}(X_t^0(\theta)) =& -\frac{1}{2}\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle + \langle \kappa^3 a_t^2, (X_t^0(\theta) - \theta^*)^{\otimes 2}\rangle \\ =& \frac{1}{2}\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle \\ \varphi_t^{\mathrm{GF}}(X_t^0(\theta)) =& \frac{1}{2}\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle + \frac{1}{2}\sigma_\varepsilon^2\langle \kappa^2, a_t^2\rangle. \end{aligned}$$

In conclusion,

$$\begin{aligned} \int_0^T \varphi_t(X_t^0(\theta))\, dt =& -\frac{1}{2}T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle \\ \int_0^T \varphi_t^{\mathrm{CC}}(X_t^0(\theta))\, dt =& \frac{1}{2}T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle \\ \int_0^T \varphi_t^{\mathrm{GF}}(X_t^0(\theta))\, dt =& \frac{1}{2}T\langle \kappa^3 e^{-2T\kappa}, (\theta - \theta^*)^{\otimes 2}\rangle + \frac{1}{2}\sigma_\varepsilon^2\langle \kappa^2, \int_0^T a_t^2\, dt\rangle. \end{aligned}$$

■

Using Proposition 4, we can compare the error terms as follows.

$$\begin{aligned}&|\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X_T^{\mathrm{CC},h})| - |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X_T^h)|\\ =&0 + \mathcal{O}(h^2), \quad \mathcal{H} \ni h \downarrow 0.\end{aligned}$$

Hence, there is no difference in the leading error term between the constant and non-constant covariance SGF approximations to (16). This justifies using the simpler CC approximation for linear regression models with Gaussian features. Moreover, we have

$$\begin{aligned}&|\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X_T^0)| - |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X_T^h)|\\ =&\frac{h}{2}\sigma_\varepsilon^2\langle\kappa^2, \int_0^T a_t^2\, dt\rangle + \mathcal{O}(h^2), \quad \mathcal{H} \ni h \downarrow 0.\end{aligned}$$

The same holds if we replace X^h by $X^{\mathrm{CC},h}$. We see that the gradient flow approximation is always worse than the approximation by a SGF due to neglecting the presence of the residuals $\boldsymbol{\varepsilon}$.

These findings are also empirically corroborated in the subsequent section for $d = 1$ (see Figure 1, lower left panel, below).

3.3 Theoretical comparison for non-Gaussian features

In this section we demonstrate that the conclusion of NCC-SGF and CC-SGF being equally good hinges on the assumption of Gaussian features.

Consider once more (16), for simplicity with $d = 1$. This time we assume merely that $\mathbb{E}\boldsymbol{x}^4 < \infty$ and $\kappa > 0$, but not that $\boldsymbol{x}$ is Gaussian. Note that

$$\begin{aligned}\mathbb{E}[(\partial_\theta \ell(\theta\boldsymbol{x}, \boldsymbol{y}))^2] =&\mathbb{E}[\boldsymbol{x}^4(\theta - \theta^*)^2 - 2\boldsymbol{\varepsilon}(\theta - \theta^*)\boldsymbol{x}^3 + \boldsymbol{x}^2\boldsymbol{\varepsilon}^2]\\ =&\kappa^2 \operatorname{Kurt}(\boldsymbol{x})(\theta - \theta^*)^2 + \kappa\sigma_\varepsilon^2,\end{aligned}$$

where $\operatorname{Kurt}(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{x}^4]/\kappa^2$ is the *kurtosis* of $\boldsymbol{x}$ (cf. section 5.1 in the appendix for more information about kurtosis). Hence,

$$\Sigma(\theta) = \operatorname{Var}[\partial_\theta \ell(\theta\boldsymbol{x}, \boldsymbol{y})] =\kappa^2(\operatorname{Kurt}\boldsymbol{x} - 1)(\theta - \theta^*)^2 + \kappa\sigma_\varepsilon^2,$$

and so the continuous-time approximations to SGD take the form

$$\begin{aligned}dX_t^0 =& -\kappa(X_t^0 - \theta^*)\, dt,\\ dX_t^h =& -\kappa(X_t^h - \theta^*)\, dt + \sqrt{h}\sqrt{\kappa^2(\operatorname{Kurt}\boldsymbol{x} - 1)(\theta - \theta^*)^2 + \kappa\sigma_\varepsilon^2}\, dW_t \qquad (19)\\ dX_t^{\mathrm{CC},h} =& -\kappa(X_t^{\mathrm{CC},h} - \theta^*)\, dt + \sqrt{h\kappa\sigma_\varepsilon^2}\, dW_t. \qquad (20)\end{aligned}$$

In analogy to Proposition 4 we can prove the following leading error expansions.

Proposition 6 *Suppose $\chi_0^h = X_0^h = X_0^{\mathrm{CC},h} = \theta \in \mathbb{R}^d$ for all $h \in (0,1)$. Then, we have*

$$\begin{aligned}
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^0) =& \frac{h}{2} T(\operatorname{Kurt} \boldsymbol{x} - 2)\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T} \\
&+ \frac{1}{4}\sigma_\varepsilon^2 \kappa(1-e^{-2\kappa T}) + \mathcal{O}(h^2) \\
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^h) =& -\frac{h}{2} T\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T} + \mathcal{O}(h^2), \\
\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^{\mathrm{CC},h}) =& \frac{h}{2} T(\operatorname{Kurt} \boldsymbol{x} - 2)\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T} + \mathcal{O}(h^2),
\end{aligned}$$

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

Proof The proof is analogous to the one of Proposition 4. Note that

$$\begin{aligned}
\varphi =& \frac{1}{2} v_t''(\mathcal{R}')^2 - \partial_t v_t' \mathcal{R}' + \frac{1}{2}\partial_t^2 v_t, \\
\varphi^{\mathrm{CC}} =& \varphi + \frac{1}{2} v_t''(\Sigma - \Sigma(\theta^*)) \\
\varphi^{\mathrm{GF}} =& \varphi^{\mathrm{CC}} + \frac{1}{2} v_t'' \Sigma(\theta^*).
\end{aligned}$$

The function φ remains the same as in the Gaussian case, while on the other hand φ^{CC} and φ^{GF} differ, since

$$v_t''(\theta)(\Sigma(\theta) - \Sigma(\theta^*)) = \kappa^3(\operatorname{Kurt} \boldsymbol{x} - 1)(\theta-\theta^*)^2 a_t^2$$

Thus,

$$\begin{aligned}
\varphi_t(X_t^0(\theta)) =& -\frac{1}{2}\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T} \\
\varphi_t^{\mathrm{CC}}(X_t^0(\theta)) =& \frac{1}{2}\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T}(\operatorname{Kurt} \boldsymbol{x} - 2) \\
\varphi_t^{\mathrm{GF}}(X_t^0(\theta)) =& \varphi_t^{\mathrm{CC}}(X_t^0(\theta)) + \frac{1}{2}\sigma_\varepsilon^2 \kappa^2 a_t^2
\end{aligned}$$

and finally

$$\begin{aligned}
\int_0^T \varphi_t(X_t^0(\theta))\, dt =& -\frac{1}{2} T\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T} \\
\int_0^T \varphi_t^{\mathrm{CC}}(X_t^0(\theta))\, dt =& \frac{1}{2} T(\operatorname{Kurt} \boldsymbol{x} - 2)\kappa^3(\theta-\theta^*)^2 e^{-2\kappa T} \\
\int_0^T \varphi_t^{\mathrm{GF}}(X_t^0(\theta))\, dt =& \int_0^T \varphi_t^{\mathrm{CC}}(X_t^0(\theta))\, dt + \frac{1}{4}\sigma_\varepsilon^2 \kappa(1-e^{-2\kappa T}).
\end{aligned}$$

■

Applying Proposition 6 to the NCC and CC-SGF approximations yields

$$\begin{aligned}
&|\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^{\mathrm{CC},h})| - |\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(X_T^h)| \\
=& \frac{h}{2}\kappa^3(\theta-\theta^*) T e^{-2\kappa T}(|\operatorname{Kurt} \boldsymbol{x} - 2| - 1) + \mathcal{O}(h^2). \qquad (21)
\end{aligned}$$

The theoretical result implies the following. For mesokurtic x, that is $\operatorname{Kurt} \boldsymbol{x} = 3$, the first summand in (21) vanishes, so there is essentially no difference in the approximation quality between NCC- and CC-SGF. The same is true for the extremal platykurtic case $\operatorname{Kurt} \boldsymbol{x} = 1$ for the simple reason that in this case the dynamics (19) and (20) coincide, hence $X^h = X^{\mathrm{CC},h}$.

For leptokurtic features, that is if $\operatorname{Kurt} \boldsymbol{x} > 3$, the first summand in (21) is positive and hence NCC is better than CC.

Among platykurtic distributions with $1 < \operatorname{Kurt} \boldsymbol{x} < 3$ we see, surprisingly, that the first summand in (21) is negative and hence the CC-SGF approximation is better.

We believe the difference between the platykurtic and leptokurtic settings may be explained as follows.

To do so, we consider 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

$$dX_t^{2,h} = -\mathcal{R}'(X_t^{2,h}) - \frac{h}{2}\mathcal{R}''(X_t^{2,h})\mathcal{R}'(X_t^{2,h})\,dt + \sqrt{h\Sigma(X_t^{2,h})}\,dW_t, \tag{22}$$

with $X_0^{2,h} = \chi_0$. Then the following holds: for every $T > 0$ there exists a $C > 0$, such that for all $g \in G^\infty(\mathbb{R})$ we have (cf. 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 first-order error term is 0. In this sense SGF2 is the best approximation we have seen so far. To achieve this improvement we have to make the drift coefficient more complicated compared to the NCC-SGF approximation. However, the SGF2 approximation has the same diffusion coefficient Σ as the NCC-SGF approximation, revealing that the diffusion part in the NCC-SGF approximation does not contribute to the leading error term.

For large $\operatorname{Kurt} \boldsymbol{x}$, Σ is large and so using an incorrect diffusion coefficient results in a large error. Hence, NCC-SGF is better then CC-SGF in this case.

On the other hand note that the leading error term for CC-SGF is the integral $\int_0^T \varphi_t^{\mathrm{CC}}(X_t^0)\,dt$, where

$$\varphi^{\mathrm{CC}} = \frac{1}{2}v_t''(\Sigma - \Sigma(\theta^*)) + \frac{1}{2}v_t''(\mathcal{R}')^2 - \partial_t v_t'\mathcal{R}' + \frac{1}{2}\partial_t^2 v_t. \tag{24}$$

The functions v and $\mathcal{R}$ do not depend on the gradient noise, that is Σ. By varying $\operatorname{Kurt} \boldsymbol{x}$ we can change the magnitude of the first summand on the RHS of (24) without affecting the other three summands, which only depend on the drift coefficient. Now, both the drift and the diffusion coefficient contribute to the leading error term. It is possible for both contributions to cancel each other out due to having opposite signs. As it turns out this happens exactly for platykurtic features. In contrast to this, in the NCC-SGF approximation only the drift coefficient contributes to the leading error term. Hence, there is no possibility for a similar cancellation.

Which SGF approximation is better in what case is summarized in Figure 2 below (cf. subsection 3.4). Moreover, we remark that for $\operatorname{Kurt} \boldsymbol{x} = 2$ the first-order error of the CC

approximation vanishes entirely. We discuss this surprising phenomenon in a numerical example in the next subsection.

Let us now end this discussion by comparing gradient flow with its stochastic versions. Comparing GF and CC-SGF, we get

$$\begin{aligned}&|\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^0_T)| - |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^{\mathrm{CC},h}_T)|\\ =&\frac{h}{4}\sigma_\varepsilon^2(1-e^{-2\kappa T}) + \mathcal{O}(h^2), \quad \mathcal{H} \ni h \downarrow 0.\end{aligned}$$

We see that the gradient flow approximation is always worse due to neglect of the noise term $\boldsymbol{\varepsilon}$. Moreover, the gap widens by increasing T.

Finally, the comparison of GF and NCC-SGF yields

$$\begin{aligned}&|\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^0_T)| - |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^h_T)|\\ =&\frac{h}{2}\kappa^3(\theta-\theta^*)Te^{-2\kappa T}(|\operatorname{Kurt}\boldsymbol{x} - 2| - 1) + \frac{h}{4}\sigma_\varepsilon^2(1-e^{-2\kappa T}) + \mathcal{O}(h^2),\end{aligned}$$

as $\mathcal{H} \ni h \downarrow 0$. Here, the situation is more complicated. For $\operatorname{Kurt}\boldsymbol{x} \approx 2$ and $\sigma_\varepsilon^2 \approx 0$ it is indeed possible that GF is better. However, for non-negligible noise terms $\boldsymbol{\varepsilon}$ or leptokurtic features NCC-SGF is clearly better.

In the next subsection will investigate and compare the weak error terms in a numerical example.

3.4 A numerical example

In this subsection we present results from a numerical experiment confirming the theoretical results presented in the previous subsections.

3.4.1 Experimental setup

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

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

with $\boldsymbol{x}, \boldsymbol{\varepsilon}$ independent, centered and of variance 1, where $\boldsymbol{\varepsilon}$ is Gaussian. We compare the weak errors of the population risk $\mathcal{R}$ for different continuous-time approximations of SGD, where $\operatorname{Kurt}\boldsymbol{x} = 1, 2, 3, 9$. To realize these kurtosises we consider, in the following order,

$$\begin{gathered}\boldsymbol{x} + \frac{1}{2} \sim \operatorname{Bin}\left(1, \frac{1}{2}\right), \quad \boldsymbol{x} + \frac{5+\sqrt{5}}{10} \sim \operatorname{Bin}\left(1, \frac{5+\sqrt{5}}{10}\right)\\ \boldsymbol{x} \sim \mathcal{N}(0,1), \quad \boldsymbol{x} + 1 \sim \operatorname{Exp}(1).\end{gathered}$$

We use a Monte Carlo approximation to estimate $\mathbb{E}\mathcal{R}(\chi^h_{T/h})$, that is

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

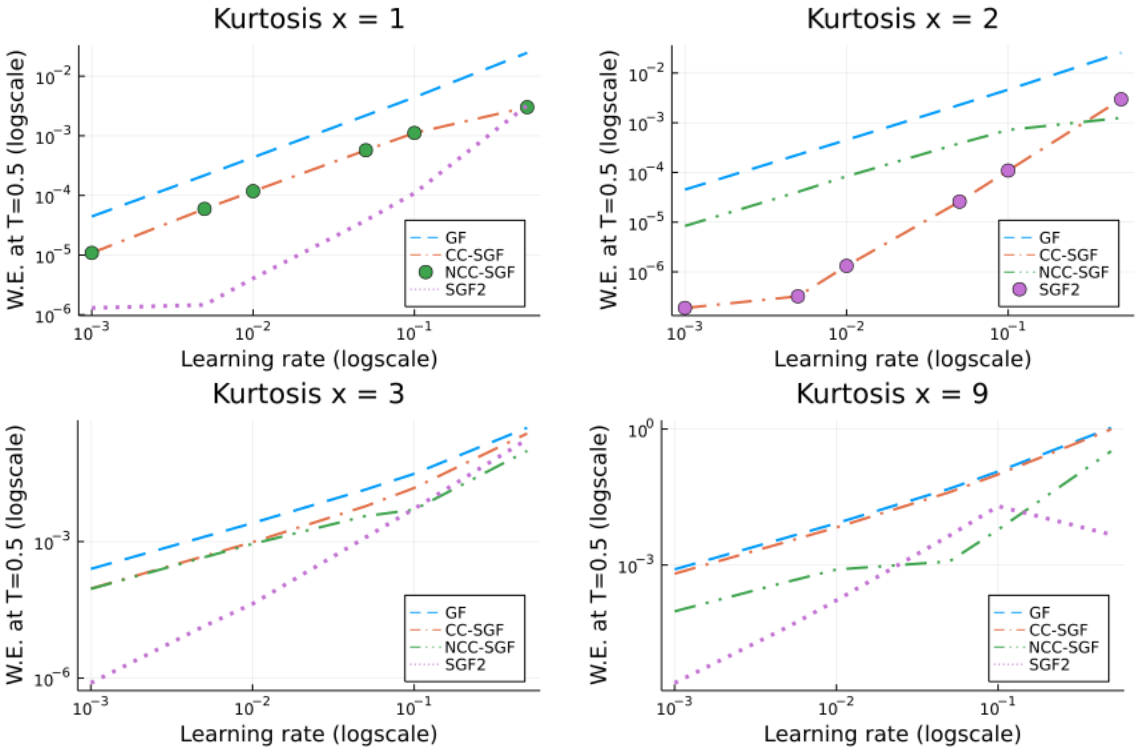

Figure 1: The weak error's dependence on the learning rate for several continuous-time approximations to SGD, for different kurtosis of the features.

where $\hat{\chi}^1, \dots, \hat{\chi}^M$ are independent copies of χ. For the experiments we have chosen M large enough (between 10^8 and 10^9) so that the variance of the Monte Carlo estimator is negligible compared to the weak error. Moreover, we determine $\mathbb{E}\mathcal{R}(Y_T^h)$ for $Y = X^0, X^h, X^{\mathrm{CC},h}, X^{2,h}$ using explicit formula, which can be derived in this example. We set $T = 0.5$ and 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. Plotted is the dependence of the weak error

$$|\mathbb{E}\mathcal{R}(\chi_{T/h}^h) - \mathbb{E}\mathcal{R}(Y_T^h)|$$

with respect to h.

3.4.2 Results

Figure 1 depicts the weak error's dependence on the learning rate for Kurt $x = 1, 2, 3, 9$.

Aside from minor deviations stemming from the Monte Carlo estimation, the empirical results in Figure 1 confirm the theoretical results in the last subsection. In particular, we observe:

(i) For a given learning rate GF has always the highest weak error, irrespective of the kurtosis. Thus, GF is always the worst approximation.

(ii) For $\operatorname{Kurt} \boldsymbol{x} = 9$ (leptokurtic) NCC- is better than CC-SGF (see Figure 1, lower right panel)

(iii) NCC- and CC-SGF are equally good for $\operatorname{Kurt} \boldsymbol{x} \in \{1, 3\}$ (see Figure 1, left panels)

(iv) For $\operatorname{Kurt} \boldsymbol{x} = 2$ the CC-SGF approximation is of second order[3] and in particular better than NCC-SGF (see Figure 1, upper right panel)

(v) The SGF2 approximation is always best, irrespective of the kurtosis.

We remark that the theoretical rates of convergence are difficult to observe without using a high number of Monte Carlo samples. Moreover, for x Bernoulli distributed ($\operatorname{Kurt} \boldsymbol{x} \in \{1, 2\}$) and $h \leq 0.001$, it is computationally difficult to get an accurate estimation of the weak error for SGF2 and for CC-SGF in the case $\operatorname{Kurt} \boldsymbol{x} = 2$. If we ignore the left-most point, then the rates are approximately 2, as predicted.

Why SGF2 and SGF can coincide The coincidence of SGF2 and CC-SGF in Figure 1 for kurtosis 2 may surprise. It can be explained as follows in this example. Equations (14) and (22) become

$$\begin{aligned} dX_t^{\mathrm{CC},h} &= -\kappa(X_t^{\mathrm{CC},h} + 1)\, dt + \sqrt{h\kappa}\, dW_t, \quad X_0^{\mathrm{CC},h} = 0, \\ dX_t^{2,h} &= -\kappa \left(1 + \frac{h}{2}\right)(X_t^{h,2} + 1)\, dt \\ &\quad + \sqrt{h}\sqrt{\kappa^2(\operatorname{Kurt} \boldsymbol{x} - 1)(X_t^{h,2} + 1)^2 + \kappa}\, dW_t, \quad X_0^{2,h} = 0. \end{aligned}$$

One can then show by direct computation that

$$\begin{aligned} \mathbb{E}[(X_t^{\mathrm{CC},h} + 1)^2] &= e^{-2\kappa t} + \frac{h}{2}(1 - e^{-2\kappa t}) \\ \mathbb{E}[(X_t^{2,h} + 1)^2] &= \frac{(h + \xi_h)e^{\xi_h \kappa t} - h}{\xi_h}, \end{aligned}$$

where $\xi_h = h\kappa(\operatorname{Kurt} \boldsymbol{x} - 2) - 2$. Moreover, using a Taylor expansion in h, we get

$$\mathbb{E}[(X_t^{2,h} + 1)^2] = e^{-2\kappa t} + \frac{h}{2}(1 - e^{-2\kappa t} + 2\kappa^2(\operatorname{Kurt} \boldsymbol{x} - 2)te^{-2\kappa t}) + \mathcal{O}(h^2), \quad h \downarrow 0.$$

Note that the term $2\kappa^2(\operatorname{Kurt} \boldsymbol{x} - 2)te^{-2\kappa t})$ vanishes exactly when $\operatorname{Kurt} \boldsymbol{x} = 2$, and in this case

$$\mathbb{E}[(X_t^{\mathrm{CC},h} + 1)^2] - \mathbb{E}[(X_t^{2,h} + 1)^2] = \mathcal{O}(h^2), \quad h \downarrow 0,$$

that is (after multiplying with $\frac{1}{2}\kappa$) we see that $\mathbb{E}\mathcal{R}(X^{\mathrm{CC},h}) - \mathbb{E}\mathcal{R}(X^{2,h})$ is on the order of h^2.

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

The difference between NCC and CC visualized In a second experiment we have analyzed how the difference of the weak errors

$$\Delta E := |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^h_T)| - |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^{\mathrm{CC},h}_T)|$$

of the two SGF approximations depends on the kurtosis of the features. To this end we consider

$$\boldsymbol{x} + p \sim \mathrm{Bin}(1, p), \quad p \in [1/2, 1]$$

and fix the learning rate to $h = 0.005$. Note that

$$\mathrm{Kurt}\,\boldsymbol{x} = K \Leftrightarrow p = \frac{K + 3 + \sqrt{K^2 + 2K - 3}}{2(K+3)}.$$

Figure 2 depicts ΔE for various values of $\mathrm{Kurt}\,\boldsymbol{x}$. We observe once more that for platykurtic features the CC approximation is better than the NCC approximation and that for leptokurtic features it is the other way around. Furthermore, both approximations are equally good for kurtosis 1 and 3. Finally, the CC approximation is best for kurtosis 2.

4 Proof of Theorems 1, 2 and 3

In this section we give proofs of our general results on leading error terms. However, before doing that we need to establish a few preliminaries.

4.1 Preliminaries

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

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

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

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

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

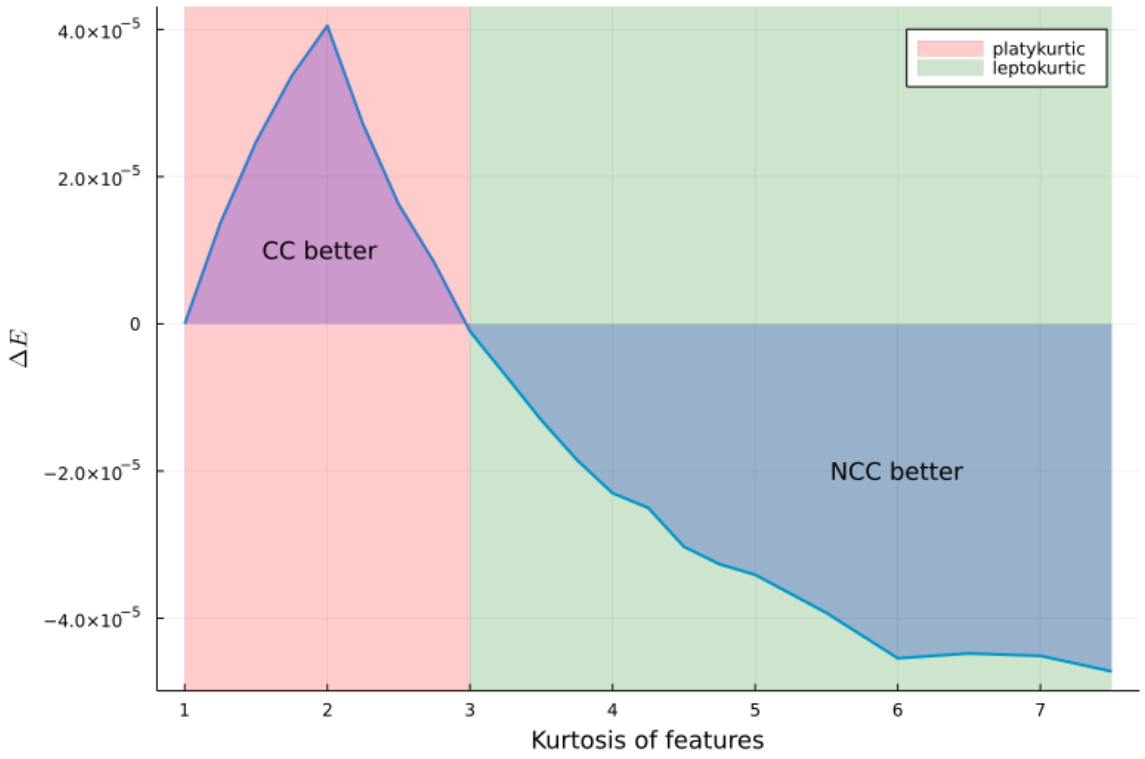

Figure 2: The empirical difference in errors $\Delta E := |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^h_T)| - |\mathbb{E}\mathcal{R}(\chi^h_{T/h}) - \mathbb{E}\mathcal{R}(X^{\mathrm{CC},h}_T)|$ as a function of kurtosis, for Bernoulli features with success probability $p \in [1/2, 1]$.

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

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

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

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

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

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

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 $n \in \mathbb{N}_0$,

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

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 $n \in \mathbb{N}_0$,

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

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,

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

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

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

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

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

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

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

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 ranges 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 22 in the Appendix. By the higher chain rule,

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

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 22 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 $U : \mathrm{Im}\, u \to \mathbb{R}$ be, 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

$$du_t = U(u_t)\,dt.$$

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[4]

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

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).

4. See for example 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 stochastic differential equation

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

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

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

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

Proof We have

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

On the one hand

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

and $x \mapsto \|b(X(x))^*\|_p \in G(\mathbb{R}^d)$ by Theorem 21 and since $b \in G_1(\mathbb{R}^d)$. On the other hand,

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

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 are $\mathbb{R}^d$-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

$$\mathbb{E}[\Delta X_s^h] = hb_s^0 + h^2 C, \qquad \mathbb{E}[(\Delta X_s^h)^{\otimes 2}] = h^2 C, \tag{28}$$

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

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

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 (cf. Theorem 23)

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

We have

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

Moreover, by Lemma 24, $\mathcal{A}_X^2 m_j \in G([0,T]\times\mathbb{R}^d)$, so Theorem 21 implies

$$\left\|(\mathcal{A}_X^2 m_j(s, X_s))\right\|_1 \leq C(1 + \||X_s|^\kappa\|_1) \leq C(1+|x|^\kappa),$$

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 24 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 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 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 $h \in \mathcal{H}$*

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

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 $t = kh$, $\delta = \Delta\chi_k^h$ 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 $\#\beta = 3 - k$,

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

since $u \leq 1$ and

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

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 $(\varphi_{nh}^{\mathrm{GF}})_{h\in\mathcal{H}, n\leq T/h}$.

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

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

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*

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

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,

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

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 leading 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

$$\left| \sum_{n=0}^{\frac{T}{h}-1} h\varphi^{\mathrm{GF}}_{nh}(X^0_{nh}) - \int_0^T \varphi^{\mathrm{GF}}_t(X^0_t)\,dt \right| \leq hC'.$$

Hence,

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

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 of (12) on $h \in \mathcal{H}$. This carries over to v and by extension to the function

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

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

$$\mathbb{E}g(\chi^h_{T/h}) - \mathbb{E}g(X^h_T) = h\int_0^T \mathbb{E}\varphi^h_t(X^h_t)\,dt + \mathcal{O}(h^2). \tag{32}$$

We then need to improve the estimate to

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

This requires an additional estimation of the difference $\varphi^h_t(X^h_t) - \varphi^0_t(X^0_t)$. 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 $x \in \mathbb{R}^d$,

$$v^h_t(x) := \mathbb{E}g(X^{h,t}_T(x)),$$

where $X^{h,t}(x)$ denotes the solution of (12) on $[t,T]$ with initial condition $X^{h,t}_t(x) = x$. 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 $(f^h_t)_{h\in(0,1),t\geq 0}$ of continuous-time stochastic processes (or merely functions) we define for every $h \in (0,1)$ the discrete-time process

$$\tilde{f}^h_n := f^h_{nh}, n \in \mathbb{N}.$$

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*

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

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*

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

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

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

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 $A_t^h := v_{t+h}^0(X_{t+h}^{h,s}) - v_t^0(X_t^{h,s})$. 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 $\Delta X_t^{h,s} := X_{t+h}^{h,s} - X_t^{h,s}$. 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 (cf. Lemma 11) we see that there exists a $C \in G(\mathbb{R}^d)$ with

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

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

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

by Theorem 21 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 (cf. 22), 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

$$x \mapsto \frac{1}{h}(\mathbb{E} w_{t+(n+1)h}^0(X_{t+(n+1)h}^h(x), \partial_j X_{t+(n+1)h}^h(x,1)) \tag{35}$$

$$- \mathbb{E} w_{t+nh}^0(X_{t+nh}^h(x), \partial_j X_{t+nh}^h(x,1))) \in G(\mathbb{R}^d) \tag{36}$$

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

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

This is a Banach space when equipped with the norm

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

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 (cf. (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

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

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 $\left\|\bar{H}\right\|_{G_1}$ 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,

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

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 21 in the Appendix. ■

Lemma 20 *We have*

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

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

Here, 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 $\left\|\tilde{X}^h\right\|_{\#\beta}$ using Theorem 21.

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

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

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

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

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

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

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

for all $h \in \mathcal{H}$ by Lemma 10, Theorem 21 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,

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

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

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

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

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 Appendix

In the appendix we provide some background on kurtosis and results from stochastic analysis.

5.1 A remark on Kurtosis

The *kurtosis* of 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, it 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	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{U}[a, b]$
Kurt.	9	$3 + \frac{1}{\lambda}$	$3 + \frac{12}{n}$	3	$\frac{9}{5}$

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. Moreover, we have $\operatorname{Kurt} Z = 2$ if and only, if $p = \frac{5 \pm \sqrt{5}}{10}$. The case of kurtosis 2 will be special for the Ornstein-Uhlenbeck approximation of SGD, as we will see later.

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 $\operatorname{Kurt} Z < 3$. We will see that these terms also delineate the settings for the error expansions in the following subsection.

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).

5.2 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 21 *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[5] *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, cf. Kunita (2004) Theorem 3.1 and 3.2 for example. The extension to an index set I and from an initial value $x \in \mathbb{R}^d$ to a process φ is discussed in 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, 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 multi-index α we define

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

Theorem 22 *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*

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

5. 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 $\|\psi_\alpha^*\|_p \in G(\mathbb{R}^d)$ 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 cf. 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 $\psi_1, \dots, \psi_{d,\dots,d}$ consists 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 of Kunita (2004) Theorem 3.4. By applying it again to the SDE governing Y the result follows via induction on l. ■

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 $|(a_1, \dots, a_n)| = n$ and $|()| = 0$ be the *length* of such a tuple.

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

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

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

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

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

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

$$\int_s^\cdot X\, dW^{\alpha-} = \left(\int_s^t X\, dW^{\alpha-}\right)_{t \geq 0} \in \mathcal{H}^{(\alpha_l)}.$$

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 23 *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, $0 \leq s \leq t \leq T, x \in \mathbb{R}^d$ and let X be the unique solution to the stochastic differential equation*

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

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

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

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

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

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

Lemma 24 *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

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International conference on machine learning*, pages 233–244. PMLR, 2020.

Jing An, Jianfeng Lu, and Lexing Ying. Stochastic modified equations for the asynchronous stochastic gradient descent. *Information and Inference: A Journal of the IMA*, 9(4): 851–873, Nov 2019. ISSN 2049-8772. doi: 10.1093/imaiai/iaz030. URL `http://dx.doi.org/10.1093/imaiai/iaz030`.

Kevin P. Balanda and H. L. MacGillivray. Kurtosis: A critical review. *The American Statistician*, 42(2):111–119, 1988. ISSN 00031305. URL http://www.jstor.org/stable/2684482.

Nicholas M Boffi and Jean-Jacques 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. URL https://books.google.de/books?id=7IRpzgEACAAJ.

Haoyang Cao and Xin Guo. Approximation and convergence of gans training: an sde approach. *arXiv preprint arXiv:2006.02047*, 2020.

Peng Chen, Qi-Man Shao, and Lihu Xu. A universal probability approximation method: Markov process approach. *arXiv preprint arXiv:2011.10985*, 2020.

Yuanyuan Feng, Lei Li, and Jian-Guo Liu. Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations. 2018.

Yuanyuan Feng, Tingran Gao, Lei Li, Jian-Guo Liu, and Yulong Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *arXiv preprint arXiv:1902.00635*, 2019.

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

Carl Graham and Denis Talay. *Stochastic simulation and Monte Carlo methods: mathematical foundations of stochastic simulation*, volume 68. Springer Science & Business Media, 2013.

Haotian Gu and Xin Guo. An SDE Framework for Adversarial Training, with Convergence and Robustness Analysis, May 2021. URL http://arxiv.org/abs/2105.08037. arXiv:2105.08037 [cs, math].

Guanqiang Hu and Yushan Zhang. Runtime analysis of stochastic gradient descent. In Ali Emrouznejad and Jui-Sheng Rayson 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. doi: 10.1145/3424978.3424993. URL https://doi.org/10.1145/3424978.3424993.

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

P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 1995. URL https://books.google.de/books?id=BCvtssom1CMC.

Daniel Kunin, Javier Sagastuy-Brena, Lauren Gillespie, Eshed Margalit, Hidenori Tanaka, Surya Ganguli, and Daniel LK Yamins. Rethinking the limiting dynamics of SGD: modified loss, phase space oscillations, and anomalous diffusion. January 2022. URL https://openreview.net/forum?id=mRc_t2b3l1-.

Hiroshi 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.

Alberto Lanconelli and Christopher S. A. Lauria. A note on diffusion limits for stochastic gradient descent, October 2022. URL http://arxiv.org/abs/2210.11257. arXiv:2210.11257 [cs, math].

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

Qianxiao Li, Cheng 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.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev 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.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 34:12712–12725, 2021.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What Happens after SGD Reaches Zero Loss? –A Mathematical Framework, July 2022. URL http://arxiv.org/abs/2110.06914. arXiv:2110.06914 [cs, stat].

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms, 2022. URL https://arxiv.org/abs/2205.10287.

Stephan Mandt, Matthew D Hoffman, David M Blei, et al. Continuous-time limit of stochastic gradient descent revisited. 2015.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, jan 2017. ISSN 1532-4435.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas 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.

Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic analysis and applications*, 8(4):483–509, 1990.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. Part II: Continuous time analysis, September 2021. URL http://arxiv.org/abs/2106.02588. arXiv:2106.02588 [cs, math, stat].

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

Jiaojiao Yang, Wenqing Hu, and Chris Junchi Li. On the fast convergence of random perturbations of the gradient flow. 2020.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong 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.