HAL
open science

Approximating stochastic gradient descent with diffusions: error expansions and impact of learning rate schedules

Stefan Ankirchner, Stefan Perko

To cite this version:

Stefan Ankirchner, Stefan Perko. Approximating stochastic gradient descent with diffusions: error expansions and impact of learning rate schedules. 2021. hal-03262396v2

HAL Id: hal-03262396

https://hal.science/hal-03262396v2

Preprint submitted on 7 Oct 2021 (v2), last revised 27 Feb 2023 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Approximating stochastic gradient descent with diffusions: error expansions and impact of learning rate schedules

Stefan Ankirchner* Stefan Perko†

October 7, 2021

Abstract

Applying a stochastic gradient descent 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 it can make sense to approximate the discrete-time process with a continuous-time diffusion. We refine some results on the weak error of diffusion approximations. In particular, we explicitly compute the leading term in the error expansion of an ODE approximation with respect to a parameter h discretizing the learning rate schedule. The leading term changes if one extends the ODE with a Brownian diffusion component. Finally, we show that if the learning rate is time varying, then its rate of change needs to enter the drift coefficient in order to obtain an approximation of order 2.

Keywords. Stochastic gradient descent; diffusion; stochastic differential equation; weak approximation; learning rate schedules; Talay-Tubaro expansion.

Introduction

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

$$\chi_{n+1} = \chi_n - \eta_n \nabla f_{\gamma(n)}(\chi_n), \quad n \in \mathbb{N}_0, \tag{0.1}$$

*Institute for Mathematics, Friedrich-Schiller-University Jena, 07737 Jena, Germany. Email: s.ankirchner@uni-jena.de

†Institute for Mathematics, Friedrich-Schiller-University Jena, 07737 Jena, Germany. Email: stefan.perko@uni-jena.de

where $(f_r)_{r\in\Gamma}$ is a family of differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$, $(\eta_n)_{n\in\mathbb{N}_0}$ is a sequence of positive reals, and $(\gamma(n))_{n\in\mathbb{N}_0}$ is an i.i.d. sequence of Γ-valued random variables. We interpret $(\chi_n)_{n\in\mathbb{N}_0}$ as the sequence of estimated parameters when applying a stochastic gradient descent (SGD) method[1] for minimizing the objective $\frac{1}{|\Gamma|}\sum_{i\in\Gamma} f_i$. We refer to η_n as the learning rate in the nth step and $f_{\gamma(n)}$ as the loss due to the n-th sample of the data or mini batch. In the following we simply call χ a SGD process.

Sometimes it is convenient to approximate the discrete-time process χ with a continuous-time stochastic process in order to make tools from Stochastic Analysis available for studying its dynamics. In a series of papers, Li, Tai and E ([13], [14]) have shown that one can approximate the distribution of χ with the distribution of a processes solving a stochastic differential equation (SDE) driven by a Brownian motion. In the following we refer to solutions of such SDEs as diffusions. We remark that the approximating diffusions are also called *stochastic modified equations* (SME), e.g. in [13] and [14].

In this paper we aim at refining some results on diffusion approximations of the SGD process (0.1). In order to take into account a time varying learning rate from the outset, we assume that the progression of the learning rate can be described in terms of a continuous function $u : [0,\infty) \to (0,1]$. We may refer to u as a *learning rate schedule*. We assume that there exists a positive real h such that for all $n \in \mathbb{N}_0$ the nth step learning rate satisfies

$$\eta_n = hu_{nh}. \tag{0.2}$$

We interpret h as a discretization parameter and use it for measuring errors of diffusion approximations. One can also view h as the maximal learning rate, since u is bounded by 1.

Let (χ_n^h) be the solution of (0.1) with initial condition $\chi_0^h = x \in \mathbb{R}^d$ and with learning rates satisfying (0.2). Moreover, let (X^h), $h \in (0,1]$ be a family of diffusions with X^h satisfying an SDE of the form

$$dX_t^h = u_t b_t^h(X_t^h)dt + u_t\sqrt{h}\sigma_t(X_t^h)dW_t, \quad X_0^h = x, \tag{0.3}$$

where W represents a Brownian motion, and b and σ are some suitable coefficients. We say that the family X^h, $h \in (0,1]$, is a weak approximation of order $l \in \mathbb{N}$ if for all $T \in (0,\infty)$ and bounded smooth functions $g : \mathbb{R}^d \to \mathbb{R}$ there exists a constant C such that for all $h \in \{T/n : n \in \mathbb{N}_0\}$ we have

$$|\mathbb{E}(g(\chi_{T/h}^\eta)) - \mathbb{E}(g(X_T^\eta))| \le Ch^l. \tag{0.4}$$

[1]It should be pointed out that by requiring $\gamma(0), \gamma(1), \ldots$ to be i.i.d. we are essentially only considering SGD *with replacement* (and variants thereof) as opposed to SGD *without replacement* which is more commonly used in practice. However the analysis of the latter is more complicated, so we will not consider it here.

By choosing $b_t^h(x) = b(x) = -\mathbb{E}[\nabla f_\gamma(x)]$ and $\sigma_t(x) = 0$ in (0.3) one obtains an ODE approximation of order 1. In our first main result we provide, for this first order ODE approximation, an explicit expression of the asymptotically minimal constant C in the error estimate (0.4). We remark that the minimal constant corresponds to the leading term in a Taylor expansion of the weak error in h. Such error expansions for numerical schemes of SDEs are referred to as Talay-Tubaro expansions (see [16]).

We then consider the diffusion family (0.3) with $b(x) = -\mathbb{E}[\nabla f_\gamma(x)]$ and $\sigma_t(x) = \sigma(x)$ given by the square root of the covariance matrix of $\nabla f_\gamma(x)$. In our second main result we show that this family is again an approximation of order 1, however with a different leading error term in the weak error expansion. Thus, we confirm a conjecture proposed by Feng et al. in [5] (Remark 2.3.). From the different leading terms we draw some consequences, e.g. we provide conditions guaranteeing that batch gradient converges at worst as slow as stochastic gradient descent.

Finally, our third main result provides a second order approximation of χ^h. In particular, the result reveals that if the learning rate is time varying, then its rate of change needs to enter the drift coefficient of any approximation of order 2.

The idea to use diffusions for approximating SGD processes appears first in [13], [14] and [15]. The approximation results in [13], [14] are rigorously shown under the assumption that the learning rate is constant and hence only for diffusion families that are time-homogeneous. In contrast, we allow for a time-dependent learning rate, and thus fall back on inhomogeneous diffusions for the weak approximation of SGD processes. In [15] the authors heuristically use an Ornstein-Uhlenbeck for approximating and analyzing the SGD process.

The article [9] considers weak approximations of order 2 for SGD processes with constant learning rates. Our third main result can be seen as a generalization of [9, Theorem 1] to the case with non-constant learning rates.

The article [6] also considers diffusion approximations for SGD processes with time-dependent learning rates, assuming that the sequence of learning rates satisfies $\eta_n = \gamma(n+1)^{-\alpha}$ for some $\gamma \in (0, \infty)$ and $\alpha \in [0, 1)$. [6, 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 [6, Theorem 1]) based on a coupling technique. In contrast to [6], we provide explicit formulas for the leading error terms, we do not make a specific assumption on the learning rate schedule u, and we incorporate the rate at which the learning rate changes into the drift coefficient in order to obtain a second order approximation.

In [2] An et al. extend the diffusion approximation framework to the setting of asynchronous SGD, where estimates of the gradient ∇f_γ arrive time-delayed by a random staleness process τ. This version of SGD is important in the distributed setting, when learning is done on multiple machines at once. The resulting diffusion dynamics are a system of two SDEs describing the expected read, i.e the (conditional) expectation of the parameters delayed by the staleness, and time-difference quotients of the expected reads scaled by the discretization factor $\Delta t = \sqrt{(1-\mu)\eta}$. Here, $\frac{1}{\mu}$ is the mean of the staleness. For $h = \eta \approx 1 - \mu$ this is comparable to our setting as $\Delta t \approx \sqrt{h^2} = h$. An approximation error of order 1 is then derived in a strong sense, which is also possible for our first-order diffusion approximation as remarked there.

In [4] the authors propose a method for approximating the limiting stationary distribution of SGD processes. The method is based on a series expansion of the backward Kolmogorov equation of a second order diffusion approximation. Convergence of order 2 is shown under a convexity assumption of the cost functions guaranteeing that the gradient process is bounded.

Theorem 3.5. in [3] by Chen et al. provides an estimate of the Wasserstein-1 distance between SGD processes and diffusion approximations with constant diffusion coefficient.

In [1] Ali et al discuss an application of the first-order diffusion approximation of SGD, which they call stochastic gradient flow. The authors derive population risk bounds between gradient flow, stochastic gradient flow and ridge regression in the mini-batch, constant learning rate setting.

In [8] Hu and Zhang study mean first passage times of SGD through the lens of its first-order diffusion approximation.

In [17] Yang et al. study the time it takes for a first-order diffusion approximation of SGD to get within a specified distance to a global minimum of a Morse loss function satisfying a "strong saddle condition". They derive an asymptotic bound for the mean stopping time depending on $\ln(\eta^{-1})$ when $\eta \downarrow 0$, where η is the (constant) learning rate.

Finally, we remark that in our first two main results we provide explicit formulas for the leading terms of weak error expansions along the parameter h. The leading terms are given in terms of integrals of the ODE solution, and thus bear similarities with the formulas of the leading weak error term when approximating SDEs with an Euler or Milstein scheme (see [16]). Indeed, our first two results 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 our second main result is given with respect to a *family* of SDEs, whereas the error considered in [16] refers to a *single* SDE.

1 Main results

We let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, \dots\}$. Let $d \in \mathbb{N}$. We write $g \in C^l$ if the function g is l-times continuously differentiable on its open domain. We may extend this to closed domains, such as $[0, T] \times \mathbb{R}^d$, by requiring $g \in C^l$ in the interior and a continuous extension of g and its derivatives to $[0, T] \times \mathbb{R}^d$.

Further, we write $g \in G(D)$ if g has (at most) polynomial growth, i.e. there exists a constant $C > 0$ and $\kappa \in \mathbb{N}_0$, such that

$$|g(x)| \leq C(1 + |x|^\kappa) \tag{1.1}$$

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 $u : [0, T] \to (0, \infty)$ be a function.

Assumption (A1) *We have $u \in C^\infty$, such that u is constant or strictly decreasing, and takes values in $[0, 1]$.*

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

$$\eta_n^h = h u_{nh}, \quad n \in \mathbb{N}_0.$$

The parameter $h \in (0, 1)$ acts as discretization parameter or maximal learning rate and is essential in describing the diffusion approximation.

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

$$\chi_{n+1}^h = \chi_n^h + \eta_n^h H_{\gamma(n)}(\chi_n^h), \quad \chi_0 = x. \tag{1.2}$$

Assumption (A2) *The function H satisfies $H \in G_1(\mathbb{R}^d)$ uniformly in $r \in \Gamma$, i.e. 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 plain (online) SGD. Given a sequence of differentiable error functions $f_1, \dots, f_M : \mathbb{R}^d \to \mathbb{R}$, where M is the sample size of our data set, we set $H_{\gamma(n)}(x) := -\nabla f_{\gamma(n)}(x)$ 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

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

Here $z^{2\otimes} = zz^T$ for any $z \in \mathbb{R}^d$. These functions appear in the coefficients of various ODE and SDE approximations of SGD. 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.*

Remark 1.1. Even with Assumption 3, Assumption 2 is still necessary by itself. This is true technically, but also holds in practically relevant settings. Consider a shallow neural network with cubic activation function given by

$$y = (\theta x)^3$$

and minimization of the square loss for two data points

$$(x^1, y^1) = (1, 0), (x^2, y^2) = (-1, 0).$$

We set $\Gamma = \{1, 2\}$, $P(\gamma(n) = 1) = \mathbb{P}(\gamma(n) = 2) = \frac{1}{2}$ for all $n \in \mathbb{N}$ and choose H_r as the derivative of the square loss due to the r-th sample for $r = 1, 2$, i.e.

$$H_r(\theta) = 3((\theta x^r)^3 - y^r)\theta^2 (x^r)^3.$$

Here we ignore our usual notational conventions and denote the argument by θ instead of x. Then $H_1(\theta) = 3\theta^5 = -H_2(\theta)$ and so $H_1, H_2 \notin G_1$, but $\bar{H} = \frac{1}{2}H_1 + \frac{1}{2}H_2 = 0 \in G_1$. $\diamondsuit$

A first order ODE approximation

We first show that the solution of the ODE

$$dX_t = u_t \bar{H}(X_t)\, dt \tag{1.3}$$

is a weak first order approximation of χ. We will refer to equation (1.3) as (generalized) gradient flow.

For convenience, we will restrict the set of acceptable learning rates to

$$\mathcal{H} := \{h \in (0,1) : T/h \in \mathbb{N}\}. \tag{1.4}$$

Let $g \in G^\infty(\mathbb{R}^d)$ and fix a time horizon $T > 0$. For all $(t,x) \in [0,T] \times \mathbb{R}^d$ we define

$$y_t(x) = g(X_T^t(x)), \tag{1.5}$$

where $X^t(x)$ denotes the solution of (1.3) on $[t,T]$ with initial condition $X_t^t(x) = x$. We write y_t^g if we want to emphasize the dependence of y on g. One can show that $y \in C^\infty([0,T]\times\mathbb{R}^d)$. Moreover, the partial derivatives of y with respect to time and space have polynomial growth in the space variable, uniformly in time. Hence, $y \in G^\infty([0,T] \times \mathbb{R}^d)$ in the sense that for every $k \in \mathbb{N}_0$ and multi-index[2] $\alpha \subseteq \{1,\dots,d\}$ there exist constants $C \in (0,\infty)$ and $\kappa \in \mathbb{N}_0$ such that

$$|\partial_t^k \partial_\alpha y_t(x)| \le C(1 + |x|^\kappa), \tag{1.6}$$

for all $t \in [0,T]$ and $x \in \mathbb{R}^d$. Then, we define the function[3]

$$\varphi_t(x) = \frac{1}{2}u_t^2 \operatorname{tr}[\nabla^2 y_t(x)\bar{H}(x)^{2\otimes}] + u_t\partial_t\nabla y_t(x)^T\bar{H}(x) + \frac{1}{2}\partial_t^2 y_t(x), \tag{1.7}$$

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 1.2. *Assume (A1), (A2) and (A3). Denote by X the solution of* (1.3) *with initial condition $X_0 = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$,*

$$\mathbb{E}g(\chi_{T/h}^h) - g(X_T) = h\int_0^T \varphi_t^g(X_t) + \frac{1}{2}u_t^2 \operatorname{tr}[\nabla^2 y_t^g(X_t)\Sigma(X_t)]\,dt + \mathcal{O}(h^2), \tag{1.8}$$

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

The parts of assumption (A3) concerning $\sqrt{\Sigma}$ are superfluous for the proof of this theorem. Recall that the discretization parameter $h \in \mathcal{H}$ can also be viewed as the maximal learning rate of SGD.

[2]See the appendix before Theorem 6.2 for a definition of (unordered) multi-indices.

[3]Here, ∇ denotes the gradient ∇^2 and the Hessian matrix with respect to x and $z^{2\otimes} := zz^T \in \mathbb{R}^{d\times d}$ for all $z \in \mathbb{R}^d$.

A first order SDE approximation

For all $h \in \mathcal{H} \cup \{0\}$ we consider the SDE

$$dX_t^h = u_t \bar{H}(X_t^h)\, dt + u_t \sqrt{h\Sigma(X_t^h)}\, dW_t. \tag{1.9}$$

Notice that as $h \to 0$ the diffusion term in (1.9) vanishes and hence (1.9) becomes the ODE (1.3).

Now let $g \in G^\infty(\mathbb{R}^d)$, $T > 0$, and consider y defined in (1.5) and φ defined in (1.7).

Theorem 1.3. *Assume (A1), (A2) and (A3). For all $h \in \mathcal{H}$ denote by X^h the solution of* (1.9) *with initial condition $X_0^h = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$,*

$$\mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^h) = h \int_0^T \varphi_t^g\, dt + \mathcal{O}(h^2), \tag{1.10}$$

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

Note that the process X^0 is the same as gradient flow defined in (1.3).

A second order SDE approximation

For all $h \in (0,1)$ we consider the SDE

$$dX_t^h = \left(u_t \bar{H}(X_t^h) - \frac{1}{2} h (u_t^2 \nabla \bar{H} \bar{H} + \dot{u}_t \bar{H})(X_t^h) \right) dt + u_t \sqrt{h\Sigma(X_t^h)}\, dW_t, \tag{1.11}$$

where $\nabla g : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ denotes the Jacobian of a function $g : \mathbb{R}^d \to \mathbb{R}^d$, i.e. $(\nabla g)_{i,j} = \partial_j f_i$ for all $i, j \in \{1, \dots, d\}$. Crucially, observe the occurrence of the $\dot{u}$ term in (1.11). If u is constant, then $\dot{u}$ vanishes. Therefore, this term was not present in previous works such as [14]. To exhibit this term we use an Itô-Taylor approximation for a time-inhomogeneous SDEs (cf. Lemma 5.1).

Theorem 1.4. *Assume (A1), (A2) and (A3). For all $h \in (0,1)$ let X^h be the solution of* (1.11) *with initial condition $X_0^h = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$ and $T > 0$,*

$$\max_{n \in \{0,\dots,\lfloor T/h \rfloor\}} |\mathbb{E}g(\chi_n^h) - \mathbb{E}g(X_{nh}^h)| \in \mathcal{O}(h^2),$$

as $h \downarrow 0$.

Remark 1.5. To simplify the notation we consider u not depending on h in Theorem 1.2, 1.3 and 1.4. However, the results can be extended to u depending on h, as long as u and its partial derivatives are bounded in h. $\diamondsuit$

Consequences

Linear regression takes fewer iterations with batch gradient descent

Recall the definition of (generalized) SGD in (1.2). We now want to compare χ to the family of deterministic processes given by

$$\chi_{k+1}^{\bar{H},h} = \chi_k^{\bar{H},h} + hu_t\bar{H}(\chi_k^{\bar{H},h}).$$

We will refer to this family of processes as *(batch) gradient descent*.

Gradient descent is itself an instance of generalized SGD, such that the increments are deterministic and in particular $\mathrm{Var}(\bar{H}) = 0$. The first-order ODE and diffusion approximation of $\chi^{\bar{H}}$ coincide. Further, they coincide with the first-order ODE approximation to χ. Theorem 1.2 implies for $T > 0$ and $g \in G^\infty(\mathbb{R}^d)$,

$$g(\chi_{T/h}^{\bar{H},h}) - g(X_T^0) = h\int_0^T \varphi_t^g(X_t^0)\,dt + \mathcal{O}(h^2), \tag{1.12}$$

for all $h \in \mathcal{H}$, where X^0 refers to the solution of (1.3) or alternatively to (1.9) with $h = 0$, and φ^g is given by (1.7) with $y_t^g := g(X_t^{0,t})$. Together with an application of Theorem 1.2 to χ we get

$$\mathbb{E}g(\chi_{T/h}^h) - g(\chi_{\bar{H},T/h}^{\bar{H},h}) = \frac{h}{2}\int_0^T u_t^2\,\mathrm{tr}[\nabla^2 y_t^g(X_t^0)\Sigma(X_t^0)]\,dt + \mathcal{O}(h^2),$$

where $\Sigma = \mathbb{E}(H_{\gamma(0)} - \bar{H})^{2\otimes}$ is the variance of the SGD increments, as usual. as a consequence we have the following.

Corollary 1.6. *Assume (A1), (A2) and (A3). For all $h \in \mathcal{H}$ denote by X^h the solution of* (1.9) *with initial condition $X_0^h = x$. Let $g \in G^\infty(\mathbb{R}^d)$ be such that y_t^g is convex in a neighborhood of x, for all $t \in [0,T]$. Then for small $h \in \mathcal{H}$,*

$$\mathbb{E}g(\chi_{T/h}^h) \geq g(\chi_{T/h}^{\bar{H},h}).$$

Given $g \in G^\infty(\mathbb{R}^d)$, we can compute

$$\nabla^2 y_t(x) = \sum_{j=1}^d \nabla\partial_j X_T^{0,t}(x)\partial_j g(X_T^{0,t}(x)) + (\nabla X_T^{0,t}(x))^2\nabla^2 g(X_T^{t,0}(x)).$$

If g is a convex loss function and

$$\sum_{j=1}^d \nabla\partial_j X_T^{0,t}(x)\partial_j g(X_T^{0,t}(x)) = 0, \tag{1.13}$$

then Corollary 1.6 tells us that, at least for small learning rates, gradient descent converges at worst as fast as SGD.

In the case of a linear ODE

$$dX_t^0 = -\kappa(X_t^0 - m)\,dt,$$

for $\kappa \in \mathbb{R}^{d\times d}$ and $m \in \mathbb{R}^d$ we have[4]

$$X_T^{0,t}(x) = (x-m)e^{-\kappa(T-t)} + m,$$

for all $t \in [0,T]$ and initial values $x \in \mathbb{R}^d$. Then $\nabla\partial_j X_T^{0,t} = 0$ for $j \in \{1,\dots,d\}$ and so the condition (1.13) is satisfied.

Gradient descent is to gradient flow what SGD is to its first-order diffusion approximation

Comparing (1.12) with Theorem 1.3 we also have the following.

Corollary 1.7. *Assume (A1), (A2) and (A3). For all $h \in \mathcal{H}$ denote by X^h the solution of* (1.9) *with initial condition $X_0^h = x$. Then for all $g \in G^\infty(\mathbb{R}^d)$,*

$$g(\chi_{T/h}^{\bar{H},h}) - g(X_T^0) = \mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^h) + \mathcal{O}(h^2), \tag{1.14}$$

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

In other words the weak approximation error between SGD and its first-order diffusion approximation is essentially matched by the approximation error between gradient descent and gradient flow, which are both deterministic rather than stochastic processes.

2 Moment estimates and growth conditions

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

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

Although usually X will be $\mathbb{R}^d$-valued and then $|\cdot|$ refers to the Euclidean norm, these definitions naturally extend to $\mathbb{R}^{d_1\times\cdots\times d_r}$-valued processes as well.

[4]Given $A \in \mathbb{R}^{d\times d}$ the matrix exponential of A is denoted by e^A.

Similarly, given an I-indexed family of discrete-time stochastic processes X we define

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

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

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

2.1 Stochastic Gradient Descent

Recall the definition of χ in (1.2), 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 $h \in (0,1)$, e.g. $Y = \chi$, we write

$$\Delta Y^{h,k}_n(x) := Y^{h,k}_{n+1}(x) - Y^{h,k}_n(x), \tag{2.1}$$

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

Lemma 2.1. *We have*

$$\begin{aligned}\mathbb{E}\Delta\chi^{h,n}_n =& \eta^h_n \bar{H},\\ \mathbb{E}(\Delta\chi^{h,n}_n)^{2\otimes} =& (\eta^h_n)^2(\Sigma + \bar{H}^{2\otimes}).\end{aligned}$$

Proof. Straightforward. □

Lemma 2.2. *The following estimates hold true:*

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

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

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

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

$$\left\|\Delta\chi^{h,n}_n(x)\right\|_p \le hC(1+|x|),$$

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

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

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

$$\begin{aligned}|\chi^h_{n+1}|^p \leq& |\chi^h_n + \eta^h_n H_{\gamma(n)}(\chi^h_n)|^p \\ \leq& |\chi^h_n|^p + \sum_{i=1}^p \binom{p}{i} |\chi^h_n|^{p-i} (\eta^h_n)^i |H_{\gamma(n)}(\chi^h_n)|^i\end{aligned}$$

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

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

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,

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

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

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

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

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

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

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

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

□

2.2 Diffusion Approximations

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

$$|f_i(x)| \leq C(1+|x|^\kappa), \tag{2.2}$$

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 (2.2).

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 2.3. *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 6.2 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 6.2 we conclude $\partial_\alpha v\in G(\mathbb{R}^d)$. □

Remark 2.4. Assume now we are given an SDE with separable coefficients, specifically

$$dX_t=u_tB(X_t)\,dt+u_tS(X_t)\,dW_t,$$

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 y defined by

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

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

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 $\mathcal{A}_Z g \in G^\infty([0,T] \times \mathbb{R}^d)$, where $\mathcal{A}_Z$ is the infinitesimal generator of Z. By Proposition 2.3 then $\mathbb{E}\mathcal{A}_Z g(Z) \in G^\infty([0,T] \times \mathbb{R}^d)$. If $g \in G^\infty(\mathbb{R}^d)$, then $y_t^{i,h} := \mathbb{E} g^i(X_T^{h,t})$ satisfies the *Feynman-Kac equation*[5]

$$\partial_t y_t + L_X y_t = 0, y_T = g,$$

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

[5] cf. [7], 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

$$\tilde{X}_n^h(x) := X_{nh}^h(x), \tag{2.3}$$

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 2.5. *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

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

and $x \mapsto \|b(X(x))^*\|_p \in G(\mathbb{R}^d)$ by Theorem 6.1 and since $b \in G_1(\mathbb{R}^d)$. On the other hand,

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

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

Proposition 2.6. *Let $l \in \mathbb{N}$, $k \in \mathbb{N}_0$,*

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

and let X be the unique solution to the stochastic differential equation

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

Suppose further we are given

$$f : I \times (0,1) \times \mathbb{N} \times \mathbb{R}^d \to \mathbb{R}, (i,h,k,x) \mapsto f_k^{i,h}(x) \in G^{l+1}(\mathbb{R}^d),$$

and that there exists a function $C \in G(\mathbb{R}^d)$, such that

$$\begin{aligned}|\mathbb{E}(\Delta\chi_k^{h,k})^\alpha - \mathbb{E}(\Delta\tilde{X}_k^{h,k})^\alpha| \leq& h^{l+1}C, \#\alpha \leq l \\ \left\|\Delta\chi_k^{h,k}\right\|_p^{l+1}, \left\|\Delta\tilde{X}_k^{h,k}\right\|_p^{l+1} \leq& h^{l+1}C, p \in \{2, 2l+2\},\end{aligned}$$

for all $h \in (0,1)$ and $k \in \{0, \dots, \lfloor T/h \rfloor\}$. Then there exists a function $C \in G(\mathbb{R}^d)$, such that

$$|\mathbb{E}f_k^{i,h}(\chi_{k+1}^{h,k}) - \mathbb{E}f_k^{i,h}(\tilde{X}_{k+1}^{h,k})| \leq h^{l+1}C,$$

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

Proof. By Taylor's theorem there exists $\theta_{\Delta\chi_k^{h,k}}, \theta_{\Delta\tilde{X}_k^{k,h}} \in (0,1)$ for every $h \in (0,1)$ and k, such that

$$\begin{aligned}f_k(\chi_{k+1}^{h,k}) - f_k(\tilde{X}_{k+1}^{h,k}) =& f_k(\chi_{k+1}^{h,k}) - f_k - (f_k(\tilde{X}_{k+1}^{h,k}) - f_k) \\ =& \sum_{0<\#\alpha\leq l} \frac{1}{\alpha!}\partial_\alpha f_k \cdot ((\Delta\chi_k^{h,k})^\alpha - (\Delta\tilde{X}_k^{k,h})^\alpha) \\ &+ \sum_{\#\beta=l+1} \sum_{D\in\Delta Y_k,\Delta Z_k} \frac{1}{\beta!}\partial_\beta f_k(\cdot + \theta_D D)D^\beta\end{aligned}$$

Since $f \in G^{l+1}(\mathbb{R}^d)$ there exists a $C \in G(\mathbb{R}^d)$, such that

$$\begin{aligned}|\mathbb{E}(\partial_\beta f(x+\theta_{D^h} D^h(x))D^h(x)^\beta)| \leq & \|\partial_\beta f\|_{G_\kappa} (1+2^{\kappa-1}|x|^\kappa + 2^{\kappa-1}\|D(x)\|_2^\kappa) \\ & \cdot \|D(x)\|_{2l+2}^{l+1} \\ \leq & c(1+|x|^\kappa + C(x))h^{l+1}C(x),\end{aligned}$$

for $\#\beta = l+1$, $D \in \Delta\chi, \Delta\tilde{X}$ and some $c > 0$ and $\kappa \in \mathbb{N}$. Therefore,

$$\begin{aligned}|\mathbb{E}f(\chi_{k+1}^{h,k}(x)) - \mathbb{E}f(\tilde{X}_{k+1}^{h,k}(x))| \leq & c \sum_{0<\#\alpha\leq l} \|\partial_\alpha f\|_{G_\kappa} (1+|x|^\kappa)h^{l+1}C \\ & + c \sum_{\#\beta=l+1} \|\partial_\beta f\|_{G_\kappa} (1+|x|^\kappa + C)h^{l+1}C,\end{aligned}$$

for some $c > 0$. □

Proposition 2.7. *Let $l \in \mathbb{N}$ and $g \in G^{l+1}(\mathbb{R}^d)$. Suppose X is given as in Proposition 2.6 and for every sequence $v : I \times (0,1) \times \mathbb{R}^d \to \mathbb{R} \in G^{l+1}(\mathbb{R}^d)$ there exists a function $C \in G(\mathbb{R}^d)$, such that*

$$|\mathbb{E}v^{i,h}(\chi_{k+1}^{h,k}) - \mathbb{E}v^{i,h}(\tilde{X}_{k+1}^{h,k})| \leq h^{l+1}C,$$

for all $i \in I$, $h \in (0,1)$ and $k \in \{0,\dots,\lfloor T/h \rfloor\}$. Further, let

$$g.P_{k,n}^h(x) := \int_{\mathbb{R}^d} g(y)\, P_{k,n}^h(x,dy) = \mathbb{E}g(\tilde{X}_n^{h,k}(x)),$$

where P is the transition kernel of $(n, \tilde{X}_n^h)_n$ and suppose

$$g.P : (k,n,h,x) \mapsto g.P_{k,n}^h(x) \in G^{l+1}(\mathbb{R}^d).$$

Then there exists a function $C \in G(\mathbb{R}^d)$, such that

$$\max_{n\in\{0,\dots,\lfloor T/h \rfloor\}} |\mathbb{E}g(\chi_n^h) - \mathbb{E}g(\tilde{X}_n^h)| \leq h^l C$$

on $\mathbb{R}^d$.

Proof. We have

$$g.P : (k,n,h,x) \mapsto g.P_{k,n}^h(x) \in G^{l+1}(\mathbb{R}^d)$$

by Proposition 2.3. Given $n \in \mathbb{N}$, $\mathbb{E}g(\tilde{X}_n) - \mathbb{E}g(\chi_n)$ equals

$$\begin{aligned}
&\sum_{k=1}^{n-1}(\mathbb{E}g(\tilde{X}_n^{k-1}\chi_{k-1}) - \mathbb{E}g(\tilde{X}_n^k\chi_k)) + \mathbb{E}g(\tilde{X}_n^{n-1}\chi_{n-1}) - \mathbb{E}g(\chi_n) \\
=&\sum_{k=1}^{n-1}\mathbb{E}\mathbb{E}(g(\tilde{X}_n^k\tilde{X}_k^{k-1}\chi_{k-1})|\tilde{X}_k^{k-1}\chi_{k-1}) - \mathbb{E}\mathbb{E}(g(\tilde{X}_n^k\chi_k)|\chi_k) \\
&+ \mathbb{E}g.P_{n,n}(\tilde{X}_n^{n-1}\chi_{n-1}) - \mathbb{E}g.P_{n,n}(\chi_n) \\
=&\sum_{k=1}^{n}(\mathbb{E}g.P_{k,n}(\tilde{X}_k^{k-1}\chi_{k-1}) - \mathbb{E}g.P_{k,n}(\chi_k)),
\end{aligned}$$

regardless of initial value $x \in \mathbb{R}^d$ or discretization parameter $h \in (0,1)$. There exists a function $C \in G(\mathbb{R}^d)$, such that

$$|\mathbb{E}g.P_{k,n}^h(\chi_{k+1}^{h,k}) - \mathbb{E}g.P_{k,n}^h(\tilde{X}_{k+1}^{h,k})| \leq h^{l+1}C,$$

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

$$|\mathbb{E}g(\tilde{X}_n^h) - \mathbb{E}g(\chi_n^h)| \leq \sum_{k=1}^{\lfloor \frac{T}{h} \rfloor} h^{l+1}\mathbb{E}C(\chi_{k-1}^h) \leq h^l TC',$$

by Lemma 2.2, for some $C' \in G(\mathbb{R}^d)$, all $h \in (0,1)$ and $n \in \{0,\dots,\lfloor T/h \rfloor\}$, since

$$\begin{aligned}
\mathbb{E}C(\chi_{k-1}^h) \leq& \|C\|_{G_\kappa}(1 + \mathbb{E}|\chi_{k-1}^h|^\kappa) \leq \|C\|_{G_\kappa}\left(1 + \sup_{h\in(0,1)}\|\chi^*\|_{\kappa,\lfloor T/h \rfloor}^\kappa\right) \\
\leq& c(1 + |\chi_0|^\kappa),
\end{aligned}$$

for some $c > 0, \kappa \in \mathbb{N}$, all $h \in (0,1)$ and $k \in \{0,\dots,\lfloor T/h \rfloor\}$. □

3 Proof of the ODE approximation

We shall give a proof of Theorem 1.2. Fix $g \in G^\infty(\mathbb{R}^d)$ and define once more $y_t(x) := g(X_T^t(x))$, where X is the solution to the gradient flow equation (1.3),

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

We then have $y \in G^\infty([0,T] \times \mathbb{R}^d)$ by Proposition 2.3 and Remark 2.4 and since we have $\bar{H} \in G^\infty(\mathbb{R}^d)$ by Assumption (A3). Further, y satisfies the *Feynman-Kac equation*

$$\partial_t y_t(x) + \nabla y_t(x)^T u_t\bar{H}(x) = 0, \quad y_T(x) = g(x). \tag{3.1}$$

From now on let χ^h and X denote the solutions of (1.2) and (1.3), respectively, with the same fixed initial condition $\chi_0 \in \mathbb{R}^d$.

Recall the definition of φ in (1.7) and the statement of Theorem 1.2. We define

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

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

Lemma 3.1. *Let $\xi : \mathcal{H} \to \mathbb{R}$ be the function such that for all $h \in \mathcal{H}$*

$$\mathbb{E} g(\chi^h_{T/h}) - g(X_T) = h^2 \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E} \Phi_{kh}(\chi^h_k) + h^2 \xi(h).$$

Then ξ is bounded.

Proof. By Taylor's theorem,

$$\begin{aligned} y_{t+h}(x+\delta) - y_t(x) =& h \partial_t y_t(x) + \sum_{j=1}^d \partial_j y_t(x) \delta_j + \frac{h^2}{2} \partial_t^2 y_t(x) \\ &+ h \sum_{j=1}^d \partial_{t,j} y_t(x) \delta_j + \frac{1}{2} \sum_{i,j}^d \partial_{i,j} y_t(x) \delta_i \delta_j \\ &+ R^h(\delta) \\ =& h \partial_t y_t(x) + \nabla y_t(x)^T \delta + \frac{h^2}{2} \partial_t^2 y_t(x) \\ &+ h \partial_t \nabla y_t(x)^T \delta + \frac{1}{2} \operatorname{tr}(\nabla^2 y_t(x) \delta^{2\otimes}) \\ &+ R^h(\delta), \end{aligned}$$

where

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

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

$$\mathbb{E} y_{(k+1)h}(\chi^h_{k+1}) - \mathbb{E} y_{kh}(\chi^h_k) = h A^h_1 + h^2 (A^h_2 + A^h_3 + A^h_4) + \mathbb{E} R^h(\Delta \chi^h_k),$$

where

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

Using (3.1) we can simplify

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

and further

$$
A_2^h = \frac{1}{2}u_{kh}^2 \mathbb{E}\operatorname{tr}[\nabla^2 y_{kh}(\chi_k^h)(\bar{H}^{2\otimes} + \Sigma)(\chi_k^h)].
$$

Moreover, 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 $|u_t| \le 1$ and

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

by Lemma 2.2. Since $\partial_t^k \partial_\alpha^{2-k} y \in G([0,T]\times\mathbb{R}^d)$ for all $k \in \{0,1,2\}$, the remainder satisfies $\mathbb{E}R^h(\Delta\chi_n^h) = \mathcal{O}(h^3)$. Therefore,

$$
\begin{aligned}
\mathbb{E}g(\chi_{T/h}^h) - g(X_T) &= \mathbb{E}y_T(\chi_{T/h}^h) - \mathbb{E}y_0(\chi_0)\\
&= \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E}y_{(k+1)h}(\chi_{k+1}^h) - \mathbb{E}y_{kh}(\chi_k^h)\\
&= h^2 \sum_{k=0}^{\frac{T}{h}-1} \mathbb{E}\Phi_{kh}(\chi_k^h) + \mathcal{O}(h^2),
\end{aligned}
$$

for all $h \in \mathcal{H}$. □

The bound on the function ξ in Lemma 3.1 only depends on the growth of g and its derivatives as well as $\bar{H}$ and Σ. We use this fact in the next step, where we apply Lemma 3.1 to the family of functions $(\Phi_{nh})_{n\geq 0, h\in\mathcal{H}}$.

For all $n \geq 1$ and $h \in \mathcal{H}$ define $\xi_n(h)$ as the real such that

$$\mathbb{E}\Phi_{nh}(\chi_n^h) - \Phi_{nh}(X_{nh}^h) = h^2 \sum_{k=0}^{n-1} \mathbb{E}\psi_{nh,kh}(\chi_k^h) + h^2\xi_n(h) \tag{3.2}$$

with

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

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

$$|\xi_n(h)| \leq B. \tag{3.3}$$

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

Lemma 3.2. *There exists a constant $C > 0$ such that*

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

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

Proof. By (3.2) and (3.3)

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

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

Because $\partial_t^k \partial_\alpha^{2-k} y \in G([0, T] \times \mathbb{R}^d)$ for all $k \in \{0, 1, 2\}$, $g \in G(\mathbb{R}^d)$, $u \in L^\infty$ and $\bar{H}, \Sigma \in G(\mathbb{R}^d)$ we have $\Phi \in G([0, T] \times \mathbb{R}^d)$. With Lemma 2.2,

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

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

Proof of Theorem 1.2. Let $g \in G^\infty(\mathbb{R}^d)$. Then Lemma 3.1 implies

$$\mathbb{E}g(\chi^h_{T/h}) - g(X_T) = h \sum_{n=0}^{\lfloor T/h \rfloor - 1} h\mathbb{E}\Phi_{nh}(\chi^h_n) + \mathcal{O}(h^2),$$

Using Lemma 3.2,

$$\begin{aligned}\sum_{n=0}^{\frac{T}{h}-1} h\mathbb{E}\Phi_{nh}(\chi^h_n) = \int_0^T \Phi_t(X_t)\,dt + h\sum_{n=0}^{\frac{T}{h}-1} \mathbb{E}\Phi_{nh}(\chi^h_n) - \Phi_{nh}(X_{nh}) \\ + \sum_{n=0}^{\frac{T}{h}-1} h\Phi_{nh}(X_{nh}) - \int_0^T \Phi_t(X_t)\,dt,\end{aligned}$$

with

$$\begin{aligned} h\sum_{n=0}^{\frac{T}{h}-1} |\mathbb{E}\Phi_{nh}(\chi^h_n) - \Phi_{nh}(X_{nh})| \leq & hC, \\ \sum_{n=0}^{\frac{T}{h}-1} |h\Phi_{nh}(X_{nh}) - \int_0^T \Phi_t(X_t)\,dt| \leq & hC'. \end{aligned}$$

Hence,

$$\mathbb{E}g(\chi^h_{T/h}) - g(X_T) = h\int_0^T \Phi_t(X_t)\,dt + \mathcal{O}(h^2),$$

for all $h \in \mathcal{H}$. □

4 Proof of the first-order SDE approximation

The proof to Theorem 1.3 is somewhat analogous to the ODE case. The diffusion coefficient makes the Feynman-Kac formula slightly more complicated, but the proof works essentially the same way.

One notable difference however comes from the newly acquired dependence of the solution X on $h \in \mathcal{H}$. This carries over to y and by extension to the function

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

Note the absence of the Σ term compared to the ODE case. By using arguments as in Section 3, we arrive at an approximation of the form

$$\mathbb{E}g(\chi^h_{T/h}) - \mathbb{E}g(X^h_T) = h\int_0^T \mathbb{E}\varphi^h_t(X^h_t)\,dt + \mathcal{O}(h^2). \tag{4.1}$$

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

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

where $X^{h,t}(x)$ denotes the solution of (1.9) on $[t,T]$ with initial condition $X^{h,t}_t(x) = x$. Then $y \in G^\infty([0,T]\times\mathbb{R}^d)$, as defined in (2.2) with $I = \mathcal{H}$, and it satisfies the Feynman-Kac equation

$$\partial_t y_t(x) + \nabla y_t^T(x) u_t \bar{H}(x) + \frac{1}{2} h u_t^2 \operatorname{tr}(\nabla^2 y_t(x)\Sigma(x)) = 0, \quad y_T(x) = g(x). \tag{4.2}$$

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 (1.2) and (1.9), respectively, with the same fixed initial condition $\chi_0 \in \mathbb{R}^d$ and $h \in \mathcal{H}$. Then we have the following.

Lemma 4.1. *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 3.1. Setting $Y := \tilde{y}$, the Taylor expansion of y gives us

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

as before, except with

$$\begin{aligned} A^h_1 :=& \mathbb{E}(\partial_t Y^h_k(\chi^h_k) + h^{-1}\nabla Y^h_k(\chi^h_k)^T \Delta\chi^h_k + \frac{1}{2} h u^2_{kh} \operatorname{tr}(\nabla^2 Y^h_k(\chi^h_k)\Sigma(\chi^h_k))) \\ =& 0 \end{aligned}$$

by (4.2) and to compensate for the additional term

$$A_2^h := \frac{1}{2} u_{kh}^2 \mathbb{E} \operatorname{tr}(\nabla^2 Y_k^h(\chi_k^h) \bar{H}^{2\otimes}(\chi_k^h)).$$

□

Again, we could have stated Lemma 4.1 with g depending on h and t, so we may show the following.

Lemma 4.2. *With the conditions as in Lemma 4.1 we have*

$$\sum_{n=0}^{\frac{T}{h}-1} |\mathbb{E}\Phi_n^h(\chi_n^h) - \mathbb{E}\Phi_n^h(\tilde{X}_n^h)| \leq \mathcal{O}(1)$$

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

$$\mathbb{E}g(\chi_{T/h}^h) - \mathbb{E}g(X_T^h) = h \int_0^T \mathbb{E}\varphi_t^h(X_t^h)\, dt + \mathcal{O}(h^2), \tag{4.3}$$

where

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

Next we shall improve (4.3) in order to arrive at the equality in Theorem 1.3. An additional step compared to the ODE approximation is then deriving an estimate of $|\mathbb{E}\varphi_t^h(X_t^h) - \varphi_0^h(X_t^0)|$ to get rid of the dependence of the integral $\int_0^T |\mathbb{E}\varphi_t^h(X_t^h)|\, dt$ on $h \in (0,1)$. First, consider estimating the difference $y^h - y^0$ and its derivatives up to order 2.

Lemma 4.4. *Let* $y_t^h(x) = \mathbb{E}g(X_T^{h,t}(x))$. *Define the* $\mathcal{H}$*-indexed family*

$$d_t^h(x) := \frac{y_t^h(x) - y_t^0(x)}{h}.$$

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

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

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

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

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

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

with some remainder term $R : \mathcal{H} \times [0,T] \times \mathbb{R}^d \to \mathbb{R} \in G([0,T] \times \mathbb{R}^d)$ and $\Delta X_t^{h,s} := X^{h,s}_{t+h} - X^{h,s}_t$. By the Feynman-Kac formula (4.2),

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

With an Itô-Taylor expansion (e.g. by using Lemma 5.1 below) we see that there exists a $C \in G(\mathbb{R}^d)$ with

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

for all $h \in (0,1)$ and $s,t \in [0,T]$ with $s \leq t$. Since ∇y^0 and $\nabla^2 y^0$ are bounded, uniformly in space and time, we conclude

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

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

$$|y_t^h - y_{nh}^h| \leq (t - nh)|\partial_t y_t^h| + h^2 R$$

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

$$|y_t^h - y_{nh}^h| \leq Ch,$$

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

$$|y_t^h - y_t^0| \leq |y_t^h - y_{nh}^h| + |y_{nh}^h - y_{nh}^0| + |y_{nh}^0 - y_t^0| \leq Ch,$$

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

Now, consider partial derivatives of y. For $j \in \{1, \ldots, d\}$ define

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

where the derivative $Y_r := \partial_j X_r^{h,t}(x, y)$ satisfies the SDE

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

with initial condition $Y_t = y$ and

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

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

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

with

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

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

$$x \mapsto \frac{1}{h} |\mathbb{E} w_{t+(n+1)h}^0 (X_{t+(n+1)h}^h(x), \partial_j X_{t+(n+1)h}^h(x, 1)) \tag{4.4}$$

$$- \mathbb{E} w_{t+nh}^0 (X_{t+nh}^h(x), \partial_j X_{t+nh}^h(x, 1))| \in G(\mathbb{R}^d) \tag{4.5}$$

and conclude, using a telescoping sum,

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

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

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

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

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

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

Consider the linear operator

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

given by

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

We have already seen it in action. Notice for example that $\varphi_t^h(x) = \mathcal{F}_t y^h(x)$ 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 4.5. *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

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

is a continuous linear operator. The statement applies for spaces of families of functions as well (cf. (2.2)*).*

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. □

Corollary 4.6. *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,

$$|\mathbb{E}\varphi^h_t(X^h_t) - \mathbb{E}\varphi^0_t(X^h_t)| \in \mathcal{O}(h) \tag{4.6}$$

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

Proof. With d defined as in Lemma 4.4 we have

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

Now apply Lemma 4.4 and the fact that $\mathcal{F}$ maps into $G([0,T]\times\mathbb{R}^d)$. With this Inequality (4.6) follows from Theorem 6.1 in the Appendix. □

Lemma 4.7. *We have*

$$|\mathbb{E}\varphi^0_t(X^h_t) - \varphi^0_t(X^0_t)| \in \mathcal{O}(h) \tag{4.7}$$

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 3.1 and its extension in (3.2), then the proof proceeds the same way. We use the Itô-Taylor approximation Lemma 5.1 to calculate $\mathbb{E}(\Delta\tilde{X}^h_n|\tilde{X}^h_n)$ and $\mathbb{E}((\Delta\tilde{X}^h_n)^{2\otimes}|\tilde{X}^h_n)$, and estimate $\left\|\tilde{X}^h\right\|_{\#\beta}$ using Theorem 6.1.

This lets us derive the expression

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

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 $x \in \mathbb{R}^d$. Then the $\mathcal{H}\times[0,T]$-indexed family $v^{h,r}_s(x) := \mathbb{E}\varphi^0_r(X^{h,s}_r(x))$ satisfies $v \in G^\infty([0,T]\times\mathbb{R}^d)$ by an extension of Remark 2.4. So Lemma 4.5 implies

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

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

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

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 2.5, Theorem 6.1 and since $\varphi^0 \in G([0,T]\times\mathbb{R}^d)$. Similarly

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

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

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

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

Proof of Theorem 1.3. Combining inequalities (4.6) and (4.7) gives us

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

for all $h \in \mathcal{H}$. We conclude with the help of (4.3),

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

□

5 Proof of the second-order SDE approximation

In this section we provide a proof of Theorem 1.4. We start with a moment estimate of SDE increments.

Lemma 5.1. *Let $b^0, b^1, \sigma \in G_1([0,\infty) \times \mathbb{R}^d) \cap G^\infty([0,\infty) \times \mathbb{R}^d)$, such that b^0, b^1 are $\mathbb{R}^d$-valued and σ is $\mathbb{R}^{d\times d}$-valued. Let $h \in (0,1), n \in \mathbb{N}$ and consider the stochastic differential equation*

$$dX_t = (b^0_t + hb^1_t)(X^h_t)\,dt + \sqrt{h}\sigma_t(X^h_t)\,dW_t, \quad X_{nh} = x,$$

with $t \in [nh, (n+1)h]$. Then there exists a function $C \in G(\mathbb{R}^d)$, such that

$$\begin{aligned} \mathbb{E}\Delta\tilde{X}^{h,n}_n =& hb^0_{nh} + \frac{1}{2}h^2(2b^1_{nh} + (\nabla b^0 b^0)_{nh} + \dot{b}^0_{nh}) + h^3 C \\ \mathbb{E}(\Delta\tilde{X}^{h,n}_n)^{2\otimes} =& h^2\left((b^0)^{2\otimes} + \sigma^T\sigma\right)_{nh} + h^3 C, \end{aligned} \tag{5.1}$$

for all $h \in (0,1)$.

Remark 5.2. The statements in Equation (5.1) are meant as statements for all initial values $x \in \mathbb{R}^d$ of the SDE. In order to simplify notation, in this section we generally omit the initial condition and formulate statements for the flow of the stochastic differential equation, i.e. the mapping from the set of initial conditions $\mathbb{R}^d$ to the collection of random variables X_t representing the corresponding solution at time $t \in [0,T]$. ♢

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, $(\Delta\tilde{X}^{h,n}_n)^\alpha = m_\alpha(X^{nh}_{(n+1)h})$. Write

$$\mathcal{A}_X = \partial_t + \mathcal{A}_{X,0} + h\mathcal{A}_{X,1}$$

with

$$\mathcal{A}_{X,0}g := (b^0 + hb^1)^T \nabla g, \mathcal{A}_{X,1}g = \frac{1}{2}\operatorname{tr}(\sigma^T\sigma\nabla^2 g).$$

Denote by $\mathcal{A}_X^i$ the i-th fold iteration of $\mathcal{A}_X$ and observe that $\mathcal{A}_X g$ already depends on time even if g does not. An Itô-Taylor expansion implies (cf. Theorem 6.3)

$$\mathbb{E}(\Delta\tilde{X}_n^{h,n})^\alpha = \sum_{i=1}^{2}\frac{h^i}{i!}\mathcal{A}_X^i m_\alpha(nh,x) + \int_{nh}^{(k+1)h}\int_{nh}^{t}\int_{nh}^{s}\mathbb{E}\mathcal{A}_X^3 m_\alpha(u,X_u)\,dudsdt.$$

We have

$$\begin{aligned}
\mathcal{A}_X(m_j)(nh,x) =& \mathcal{A}_{X,0}(m_j)\\
=& b_{nh}^0(x)_j + hb_{nh}^1(x)_j\\
\mathcal{A}_X^2(m_j)(nh,x) =& (\mathcal{A}_{X,0}^2 + \mathcal{A}_{X,1}\mathcal{A}_{X,0} + \partial_t\mathcal{A}_{X,0})(m_j)(nh,x)\\
=& (\nabla b^0 b^0 + h(\nabla b^0 b^1 + \nabla b^1 b^0) + h^2\nabla b^1 b^1)_{nh}(x)_j\\
&+ \frac{1}{2}h\operatorname{tr}(\nabla^2 b_j^0\sigma^T\sigma)_{nh}(x) + \frac{1}{2}h^2\operatorname{tr}(\nabla^2 b_j^1\sigma^T\sigma)(x)\\
&+ \dot{b}_{nh}^0(x)_j + h\dot{b}_{nh}^1(x)_j,
\end{aligned}$$

where $\nabla^2 g_j$ is simply the Hessian of $g_j : \mathbb{R}^d \to \mathbb{R}$. Therefore,

$$\begin{aligned}
\sum_{i=0}^{2}\frac{1}{i!}h^i\mathcal{A}_X^i m_j(nh,x) =& hb_{nh}^0(x) + h^2\left(b_{nh}^1 + \frac{1}{2}(\nabla b^0 b^0)_{nh} + \frac{1}{2}\dot{b}_{nh}^0\right)(x)_j\\
&+ h^3 C(x),
\end{aligned}$$

for some $C \in G(\mathbb{R}^d)$. By Lemma 6.4, $\mathcal{A}_X^3 m_j \in G([0,T]\times\mathbb{R}^d)$ and by Theorem 6.1 we have

$$\left\|(\mathcal{A}_X^3 m_j(s,X_s))\right\|_1 \le C(1 + \||X_s|^\kappa\|_1) \le C(1+|x|^\kappa),$$

for some constant $C > 0$. Hence,

$$\begin{aligned}
\left|\int_{nh}^{(k+1)h}\int_{nh}^{t}\int_{nh}^{s}\mathbb{E}\mathcal{A}_X^3 m_j(u,X_u)\,dudsdt\right| \le& \int_{nh}^{(k+1)h}\int_{nh}^{t}\int_{nh}^{s}C(x)\,du\,ds\,dt\\
=& h^3 C'(x),
\end{aligned}$$

with $C, C' \in G(\mathbb{R}^d)$, and so

$$\mathbb{E}\Delta X_{nh}^h = hb_{nh}^0 + \frac{1}{2}h^2\left(2b_{nh}^1 + (\nabla b^0 b^0)_{nh} + \dot{b}_{nh}^0\right) + h^3 C(x),$$

for some $C \in G(\mathbb{R}^d)$.

Now, let us consider a multi-index $\alpha = \{j_1, j_2\}$. Writing $f_{nh}(x)_\alpha = [f_{nh}(x)]_{j_1,j_2}$ and $(f_\alpha)_{nh}(x) = (f_{j_1,j_2})_{nh}(x)$, we have

$$\begin{aligned}
\mathcal{A}_X(m_\alpha)(nh,x) =& \mathcal{A}_{X,1}(nh,m_\alpha) = \frac{1}{2}h(\sigma^T\sigma)_{nh}(x)_\alpha \\
\mathcal{A}_X^2(m_\alpha)(nh,x) =& (\mathcal{A}_{X,0}^2 + \mathcal{A}_{X,0}\mathcal{A}_{X,1} + \mathcal{A}_{X,1}^2 + \partial_t\mathcal{A}_{X,1})(m_\alpha)(nh,x) \\
=& (b^0(b^0)^T + h(b^0(b^1)^T + b^1(b^0)^T) + h^2 b^1(b^1)^T)_{nh}(x)_\alpha \\
&+ \frac{1}{2}\left(h\sum_{l=1}^d b_l^0\partial_l(\sigma^T\sigma)_\alpha + h^2\sum_{l=1}^d b_l^1\partial_l(\sigma^T\sigma)_\alpha\right)_{nh}(x) \\
&+ \frac{1}{4}h^2\operatorname{tr}(\nabla^2(\sigma^T\sigma)_\alpha\sigma^T\sigma)_{nh}(x) \\
&+ \frac{1}{2}h\partial_t(\sigma^T\sigma)_{nh}(x)_\alpha,
\end{aligned}$$

where $\nabla^2(\sigma^T\sigma)_\alpha$ is the Hessian of $(\sigma^T\sigma)_{j_1,j_2}$, and so again using Lemma 6.4,

$$\mathbb{E}(\Delta X_{nh}^h)^{2\otimes} = h^2\left((b^0)^{2\otimes} + \sigma^T\sigma\right)_{nh} + h^3 C,$$

for some $C \in G(\mathbb{R}^d)$. □

Remark 5.3. With Lemma 5.1 and 2.1 we may compare SGD with the solution of the family of SDE's

$$dX_t^h = (b_t^0 + hb_t^1)(X_t^h)\,dt + \sqrt{h}\sigma_t(X_t^h)\,dW_t, \quad X_{nh}^h = x,$$

with the choice $\eta_k^h = hu_{nh}$.

$$\begin{aligned}
\mathbb{E}\Delta\chi_k^h - \mathbb{E}\Delta\tilde{X}_n^{h,n} =& h(u_{nh}\bar{H} - b_{nh}^0) + \frac{1}{2}h^2(2b_{nh}^1 + (\nabla b^0 b^0)_{nh} + \dot{b}_{nh}^0) + h^3 C, \\
\mathbb{E}(\Delta\chi_k^h)^{2\otimes} - \mathbb{E}(\Delta\tilde{X}_n^{h,n})^{2\otimes} =& h^2(u_t^2\Sigma - \sigma^T\sigma + u_t^2\bar{H}^{2\otimes} - (b^0)^{2\otimes})_{nh} + h^3 C.
\end{aligned}$$

This gives us an idea of how to choose the coefficients b^0, b^1 and σ, which is

$$b_t^0 := u_t\bar{H}, \quad b_t^1 := -\frac{1}{2}\left(u_t^2\nabla\bar{H}\bar{H} + \dot{u}_t\bar{H}\right), \quad \sigma_t := u_t\sqrt{\Sigma}.$$

Note that the conditions

$$\bar{H}, \sqrt{\Sigma} \in G_1(\mathbb{R}^d), \bar{H}, \Sigma \in G^\infty(\mathbb{R}^d), u \in C^\infty([0,1]\times[0,T])$$

are enough to satisfy the assumptions of Lemma 5.1 for all $h \in (0,1)$ and $n \in \{0,\dots,\lfloor T/h\rfloor\}$. ♢

We are finally ready to prove Theorem 1.4.

Proof of the Main Theorem 1.4. By Rem. 5.3

$$|\mathbb{E}(\Delta\chi_n^{h,n})^\alpha - \mathbb{E}(\Delta\tilde{X}_n^{h,n})^\alpha| \leq h^3 C,$$

for $\#\alpha \leq 2$ and by Lemma 2.2 and 2.5

$$\left\|\Delta\chi_n^{h,n}\right\|_p^3 \vee \left\|\Delta X_n^{h,n}\right\|_p^3 \leq h^3 C$$

for all $n \in \mathbb{N}, h \in (0,1)$ and some $C \in G(\mathbb{R}^d)$. Therefore, given any $g \in G^\infty(\mathbb{R}^d)$, by Proposition 2.6 with $I = \{(i+1,n) \in \mathbb{N}^2 : i < n\}$,

$$\left|\mathbb{E}g.P^h_{k+1,n}(\chi^{h,k}_{k+1}) - \mathbb{E}g.P^h_{k+1,n}(X^{h,kh}_{(k+1)h})\right| \leq h^3 C$$

for some $C \in G(\mathbb{R}^d)$, where P^h are transition kernels of $(X^h_{nh})_{n\in\mathbb{N}_0}$. Then, by Proposition 2.7 together with Lemma 2.2 and Proposition 2.3,

$$\max_{n\in\{0,\dots,\lfloor T/h\rfloor\}} |\mathbb{E}g(X^h_{nh}) - \mathbb{E}g(\chi^h_n)| \leq h^2 C$$

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

Remark 5.4. We can improve Theorem 1.4 to include random initial values. Let $\xi \in L^2$ be independent of γ and the filtration $\mathcal{F}$. Then,

$$\begin{aligned}
&\max_{n\in\{0,\dots,\lfloor T/h\rfloor\}} |\mathbb{E}g(\chi^h_n(\xi)) - \mathbb{E}g(X^h_{nh}(\xi))| \\
=&\max_{n\in\{0,\dots,\lfloor T/h\rfloor\}} |\mathbb{E}(\mathbb{E}g(\chi^h_n(x)) - \mathbb{E}g(X^h_{nh}(x))|\xi = x)| \\
\leq&\mathbb{E}\left(\max_{n\in\{0,\dots,\lfloor T/h\rfloor\}} |\mathbb{E}g(\chi^h_n(x)) - \mathbb{E}g(X^h_{nh}(x))|\,|\xi = x\right) \\
\leq&h^2 C(\xi),
\end{aligned}$$

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

6 Appendix

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 6.1. *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[6] 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. [11] 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 [14] Theorem 18 and 19. □

A (unordered) *multi-index* $\alpha \subseteq \{1,\dots,d\}$ is a multi-subset of $\{1,\dots,d\}$, i.e. 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

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

[6] Of course, we mean unique up to indistinguishability.

Theorem 6.2. *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*

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

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. [11] Theorem 3.4. More specifically, for every $l \in \mathbb{N}$, assuming the result holds for all $l' < l$ define

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

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

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

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 [11] 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, i.e.

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

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

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

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

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

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

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

Further, applying expectation yields

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

Proof. See [10] 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 $|\alpha| \le l$. As the hierarchical set choose $\mathcal{A} := \{\alpha : |\alpha| \le l\}$. For the second statement note that

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

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

Lemma 6.4. *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 $i \le \frac{l}{2} \wedge l'$, 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,

$$\mathcal{A}_{X,0} g = \sum_{j=1}^{d} b_j \partial_j g \in G^{l-2i+1}(\mathbb{R}^d),$$

$$\mathcal{A}_{X,1} g = \sum_{j,k}^{d} (\sigma^T \sigma)_{j,k} \partial_{j,k} g \in G^{l-2i}(\mathbb{R}^d),$$

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

[1] A. Ali, E. Dobriban, and R. J. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. 2020.

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

[3] P. Chen, Q.-M. Shao, and L. Xu. A universal probability approximation method: Markov process approach. 2020.

[4] Y. Feng, T. Gao, L. Li, J.-G. Liu, and Y. Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *Commun. Math. Sci.*, 18(1):163–188, 2020.

[5] Y. Feng, L. Li, and J.-G. Liu. Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations. 2018.

[6] X. Fontaine, V. De Bortoli, and A. Durmus. Continuous and discrete-time analysis of stochastic gradient descent for convex and non-convex functions. *arXiv preprint arXiv:2004.04193*, 2020.

[7] C. Graham and D. Talay. *Stochastic Simulation and Monte Carlo Methods.* 2013.

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

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

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

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

[12] J. Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(4), May 2021.

[13] Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. 2015.

[14] Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. 2018.

[15] S. Mandt, M. D. Hoffman, D. M. Blei, et al. Continuous-time limit of stochastic gradient descent revisited. In *OPT workshop, NIPS*, 2015.

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

[17] J. Yang, W. Hu, and C. J. Li. On the fast convergence of random perturbations of the gradient flow. 2020.