A note on diffusion limits for stochastic gradient descent

Alberto Lanconelli* Christopher S. A. Lauria†

January 30, 2025

Abstract

In the machine learning literature stochastic gradient descent has recently been widely discussed for its purported implicit regularization properties. Much of the theory, that attempts to clarify the role of noise in stochastic gradient algorithms, has approximated stochastic gradient descent by a stochastic differential equation with Gaussian noise. We provide a rigorous theoretical justification for this practice that showcases how the Gaussianity of the noise arises naturally.

Key words and phrases: stochastic gradient descent, Markov process, Brownian motion, stochastic differential equation.

Mathematics Subject Classification 2020: 65K05; 60F17; 60H10.

1 Introduction

Stochastic gradient descent (SGD) ([Robbins and Monro, 1951], [Nemirovski et al., 2009]) is an optimization algorithm that has found a great deal of success in applied settings, especially in the machine learning framework [Tsypkin and Nikolic, 1971], [Bottou, 1998], [Bottou et al., 2018]. Its success has spurred the creation of many variants, [Rumelhart et al., 1986], [Kingma and Ba, 2014], [Duchi et al., 2011], and its properties have been studied in a number of ways, [Netrapalli, 2019], [Gower et al., 2019], [Nguyen et al.,], [Moulines and Bach, 2011], [Toulis and Airoldi, 2017].
A widespread setting where SGD is used concerns minimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x) := \min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{1.1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \dots, n$ represent some given *loss functions*. The SGD recursive iteration, in its simplest form, can be written as

$$x_0^{(\eta)} = x_0 \in \mathbb{R}^d \quad \text{and} \quad x_{k+1}^{(\eta)} = x_k^{(\eta)} - \eta \nabla f_{\gamma_{k+1}}(x_k^{(\eta)}), \quad k \geq 0 \tag{1.2}$$

where η is a positive constant, usually named *learning rate*, while $\{\gamma_k\}_{k\in\mathbb{N}}$ is a sequence of independent discrete uniform random variables on the set $\{1, 2, \dots, n\}$. Here we stress through the superscript η in $x_k^{(\eta)}$ the crucial dependence of the iterative scheme on the size of the learning rate.

Recently, the iteration scheme (1.2) has been approximated, in the limit as η tends to zero, by certain continuous time analogues which can be formalized in terms of stochastic differential equations

*Dipartimento di Scienze Statistiche Paolo Fortunati, Università di Bologna, Bologna, Italy. **e-mail**: alberto.lanconelli2@unibo.it

†Dipartimento di Scienze Statistiche Paolo Fortunati, Università di Bologna, Bologna, Italy. **e-mail**: christopher.lauria2@unibo.it

(SDE), see for instance [Mandt et al., 2015] and [Chaudhari and Soatto, 2018]. We refer the reader also to [Wang and Wu, 2020] where the rate of convergence to the diffusion limit is presented. These types of approximations permit an asymptotic analysis of stochastic gradient algorithms. Moreover, the continuous time treatment also allows the application of optimal control theory to study the problems of adaptive hyper-parameter adjustments [Li et al., 2017].
The SDE approximation to SGD has been used throughout the literature; the most rigorous formalization of it has been given by [Li et al., 2017] and [Li et al., 2019] (see Section 3.1), where stochastic gradient algorithms are shown to be approximated in distribution by SDEs driven by, appropriately chosen, Gaussian noise. Other authors have studied SGD in specific contexts where heavy-tailed noise is considered, see [Nguyen et al., 2019], [Simsekli et al., 2019], see also [Li et al., 2021].

The aim of this note is to provide a rigorous justification for the heuristic argument proposed in [Li et al., 2017] and [Li et al., 2019], in particular we show that the Gaussian nature of the noise arises naturally thus providing an additional theoretical justification for its pervasive use in the literature. Our approach, inspired by a similar setting in the econometric literature [Nelson, 1990] (see also [Corradi, 2000] and [Buccheri et al., 2021]), utilizes a general theorem from [Stroock and Varadhan, 1997] which provides sufficient conditions for a sequence of discrete time Markov processes to converge in distribution to an Itô SDEs driven by Brownian motion.

2 Heuristic derivation of the diffusion limit and statement of the main result

Before stating our main result we summarize the procedure proposed by the authors in [Li et al., 2017] and [Li et al., 2019] to derive an SDE from the iterative scheme (1.2). The main idea is to treat the learning rate η as the time step of a discretization scheme for a continuous time stochastic process. To this aim the iterative rule (1.2) will be more conveniently written as

$$x^{(\eta)}_{\eta(k+1)} = x^{(\eta)}_{\eta k} - \eta \nabla f_{\gamma_{k+1}}(x^{(\eta)}_{\eta k}), \tag{2.1}$$

so that the difference between the lower indexes of $x^{(\eta)}_{\eta(k+1)}$ and $x^{(\eta)}_{\eta k}$ agrees with the proportionality constant in front of the gradient $\nabla f_{\gamma_{k+1}}(x^{(\eta)}_{\eta k})$.
The two preparatory steps used in [Li et al., 2019] for relating (2.1) to an SDE are:

- Rewriting (2.1) (recalling the definition of f in (1.1)) as

$$\begin{aligned} x^{(\eta)}_{\eta(k+1)} &= x^{(\eta)}_{\eta k} - \eta \nabla f_{\gamma_{k+1}}(x^{(\eta)}_{\eta k}) \\ &= x^{(\eta)}_{\eta k} - \eta \left(\nabla f_{\gamma_{k+1}}(x^{(\eta)}_{\eta k}) - \nabla f(x^{(\eta)}_{\eta k}) + \nabla f(x^{(\eta)}_{\eta k}) \right) \\ &= x^{(\eta)}_{\eta k} - \nabla f(x^{(\eta)}_{\eta k})\eta + \left(\nabla f(x^{(\eta)}_{\eta k}) - \nabla f_{\gamma_{k+1}}(x^{(\eta)}_{\eta k}) \right) \eta \\ &= x^{(\eta)}_{\eta k} - \nabla f(x^{(\eta)}_{\eta k})\eta + V^{(\eta)}_{\eta k}\sqrt{\eta}, \end{aligned} \tag{2.2}$$

  where we set

$$V^{(\eta)}_{\eta k} := (\nabla f(x^{(\eta)}_{\eta k}) - \nabla f_{\gamma_{k+1}}(x^{(\eta)}_{\eta k}))\sqrt{\eta}, \quad k \geq 0.$$

  Notice that $V^{(\eta)}_{\eta k}$ is a d-dimensional random vector with conditional mean

$$\mathbb{E}[V^{(\eta)}_{\eta k} | \mathcal{F}^{(\eta)}_{\eta k}] = 0$$

  and conditional covariance matrix

$$\mathbb{E}\left[V^{(\eta)}_{\eta k} (V^{(\eta)}_{\eta k})^T | \mathcal{F}^{(\eta)}_{\eta k} \right] = \frac{\eta}{n} \sum_{i=1}^{n} (\nabla f(x^{(\eta)}_{\eta k}) - \nabla f_i(x^{(\eta)}_{\eta k}))(\nabla f(x^{(\eta)}_{\eta k}) - \nabla f_i(x^{(\eta)}_{\eta k}))^T$$

$$= \eta\Sigma(x_{\eta k}^{(\eta)}),$$

where we denoted

$$\Sigma(x) := \frac{1}{n}\sum_{i=1}^{n}(\nabla f(x) - \nabla f_i(x))(\nabla f(x) - \nabla f_i(x))^T, \tag{2.3}$$

while $\mathcal{F}_{\eta k}^{(\eta)}$ stands for the σ-algebra generated by the random vectors $x_0^{(\eta)}, x_\eta^{(\eta)}, ..., x_{\eta k}^{(\eta)}$ for $k \geq 0$.

- Replacing $V_{\eta k}^{(\eta)}$ with

$$\overline{V_{\eta k}^{(\eta)}} := (\nabla f(x_{\eta k}^{(\eta)}) - \nabla f_{\gamma_{k+1}}(x_{\eta k}^{(\eta)}))\sqrt{\overline{\eta}}, \quad k \geq 0,$$

where $\overline{\eta}$ plays the role of a new parameter independent of η; observe that the conditional covariance matrix of $\overline{V_{\eta k}^{(\eta)}}$ is now

$$\mathbb{E}\left[\overline{V_{\eta k}^{(\eta)}}(\overline{V_{\eta k}^{(\eta)}})^T|\mathcal{F}_{\eta k}^{(\eta)}\right] = \overline{\eta}\Sigma(x_{\eta k}^{(\eta)}) \tag{2.4}$$

(we remark that the splitting of η as $\sqrt{\eta}\sqrt{\overline{\eta}}$ is necessary for the appearance of a diffusion term in the recursive equation). With such a substitution equation (2.2) reads

$$x_{\eta(k+1)}^{(\eta)} = x_{\eta k}^{(\eta)} - \nabla f(x_{\eta k}^{(\eta)})\eta + \overline{V_{\eta k}^{(\eta)}}\sqrt{\eta}, \quad k \geq 0. \tag{2.5}$$

At this point, the authors in [Li et al., 2019] take a further step and replace the term $\overline{V_{\eta k}^{(\eta)}}\sqrt{\eta}$ in (2.5) with a Gaussian random vector having the same mean and covariance as $\overline{V_{\eta k}^{(\eta)}}\sqrt{\eta}$; more precisely, they consider the random vector

$$\left(\overline{\eta}\Sigma(x_{\eta k}^{(\eta)})\right)^{\frac{1}{2}}(B_{(k+1)\eta} - B_{k\eta}),$$

with $\{B_t\}_{t\geq 0}$ being a d-dimensional standard Brownian motion, which transforms (2.5) into

$$x_{\eta(k+1)}^{(\eta)} = x_{\eta k}^{(\eta)} - \nabla f(x_{\eta k}^{(\eta)})\eta + \left(\overline{\eta}\Sigma(x_{\eta k}^{(\eta)})\right)^{\frac{1}{2}}(B_{(k+1)\eta} - B_{k\eta}), \quad k \geq 0,$$

and we are going to rigorously justify this substitution. This equation corresponds to the Euler scheme (see for instance [Kloeden and Platen, 2011]) with step size η for the d-dimensional Itô's SDE

$$dX_t = -\nabla f(X_t)dt + (\overline{\eta}\Sigma(X_t))^{\frac{1}{2}}\,dB_t, \quad t \geq 0. \tag{2.6}$$

The SDE in (2.6) is then taken as a candidate for the continuous time limit of the iteration scheme (2.1). Subsequently, [Li et al., 2019] prove that for any $\mathtt{T} > 0$, $k \in \{0, ..., [\mathtt{T}/\bar{\eta}]\}$ and any polynomially bounded function $g : \mathbb{R}^d \to \mathbb{R}$ one has

$$|\mathbb{E}[g(X_{\bar{\eta}k})] - \mathbb{E}[g(x_k^{(\bar{\eta})})]| \leq C\bar{\eta}.$$

The solution to (2.6) is thus interpreted as a weak asymptotic approximation, as $\bar{\eta}$ tends to zero, of the SGD update $x_k^{(\bar{\eta})}$ in (1.2).

The aim of this note is to show that (2.6) can be rigorously derived from (2.5), thus entailing the intrinsic Gaussian nature of the noise. To do so we will invoke a general functional limit theorem for discrete time Markov processes established in [Stroock and Varadhan, 1997]. Our result reads as follows.

Theorem 2.1. *Assume that:*

1. *the gradients ∇f_i, $i = 1, ..., n$ are locally Lipschitz continuous;*
2. *for all $R > 0$ there exists a positive real constant λ_R such that the covariance matrix Σ defined in* (2.3) *obeys the bound*

$$\max_{j\in\{1,\dots,d\}} \left|\langle\theta, \partial^2_{x_j}\Sigma(x)\theta\rangle\right| \leq \lambda_R\|\theta\|^2, \quad \theta \in \mathbb{R}^d \tag{2.7}$$

for all $x \in \mathbb{R}^d$ such that $\|x\| < R$ (here, the differential operator $\partial^2_{x_j}$ acts on the matrix $\Sigma(x)$ entrywise).

Then, the discrete time stochastic process $\{x^{(\eta)}_{\eta k}\}_{k\geq 0}$ defined recursively in (2.5) *converges in distribution, as η tends to zero, to the unique weak solution of the Itô SDE* (2.6).

The Gaussian nature of the limiting object arises naturally in Theorem 2.1 through the application of the Stroock-Varadhan-Nelson general theorem. This distinguishes Theorem 2.1 from the convergence results in [Wang and Wu, 2020], which relies on different and stronger assumptions, including global rather than local Lipschitz continuity of the loss function, strong convexity (unnecessary in our framework), and positive definiteness of the covariance matrix (we permit degenerate cases). In addition, the proof in [Wang and Wu, 2020] is tailored on the stochastic gradient descent algorithm with an arbitrary batch size while ours follows from a general diffusion limit approximation making direct comparison hard to draw.

Remark 2.2. *Neural networks generally have a locally Lipschitz gradient with respect to their parameters, given common network architectures (e.g., feedforward, convolutional) and typical activation functions (e.g., ReLU). On the other hand the condition on the covariance matrix Σ is different from the bounded gradient noise assumption commonly used in the literature to prove the convergence of SGD. In this context, it is needed to ensure that the limiting equation* (2.6) *behaves appropriately.*

Remark 2.3. *The weak, i.e., in the sense of probability distributions, nature of the convergence proven in Theorem 2.1 is coherent with the features of our original problem. In fact, the randomness introduced in the gradient descent scheme is somewhat artificial (meaning it is not inherently linked to the problem's data) and only its law has a role in the development of the limit.*

3 Proof of Theorem 2.1

The proof of our main result consists in an application of a general functional limit theorem stated in [Stroock and Varadhan, 1997]. Here, we utilize a slightly different version, proposed in [Nelson, 1990], which we further simplify to match our framework. The next section is devoted to its rigorous statement.

3.1 A general functional limit theorem

Let $\left\{x^{(\eta)}_{\eta k}\right\}_{k\geq 0}$ be a d-dimensional discrete time Markov process defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by the positive real number η. We assume that

$$\mathbb{P}\left(x^{(\eta)}_0 = x_0\right) = 1, \quad \text{for all } \eta > 0$$

where x_0 is a deterministic d-dimensional vector. For $\eta > 0$ and $k \geq 0$ we write $\mathcal{F}^{(\eta)}_{k\eta}$ to represent the σ-algebra generated by the random vectors $x_0, x^{(\eta)}_\eta, x^{(\eta)}_{2\eta}, ..., x^{(\eta)}_{k\eta}$; the Markov property of $\left\{x^{(\eta)}_{\eta k}\right\}_{k\geq 0}$ yields

$$\mathbb{E}[\varphi(x^{(\eta)}_{\eta(k+1)})|\mathcal{F}^{(\eta)}_{k\eta}] = \mathbb{E}[\varphi(x^{(\eta)}_{\eta(k+1)})|x^{(\eta)}_{\eta k}] = \int_{\mathbb{R}^d} \varphi(y)\Pi^{(\eta)}_{\eta k}(x^{(\eta)}_{\eta k}; dy),$$

for any bounded and measurable $\varphi : \mathbb{R}^d \to \mathbb{R}$; here, $\Pi_{\eta k}^{(\eta)}$ denotes the transition kernel of the process $\left\{x_{\eta k}^{(\eta)}\right\}_{k\geq 0}$. We now embed the discrete time process $\left\{x_{\eta k}^{(\eta)}\right\}_{k\geq 0}$ in a continuous time one, denoted $\{X_t^{(\eta)}\}_{t\geq 0}$, through the prescription

$$X_t^{(\eta)} := x_{k\eta}^{(\eta)}, \quad \text{if } t \in [k\eta, (k+1)\eta[. \tag{3.1}$$

The following set of assumptions will imply the weak convergence of $\{X_t^{(\eta)}\}_{t\geq 0}$ towards the solution of an Itô SDE, as η tends to zero.

Assumption 3.1. *There exist continuous functions $a : \mathbb{R}^d \times [0,\infty[\to S_d$, the space of $d \times d$ symmetric non negative definite matrices, and $b : \mathbb{R}^d \times [0,\infty[\to \mathbb{R}^d$ such that for all $R > 0$ and $\mathtt{T} > 0$ we get*

$$\lim_{\eta\to 0} \sup_{\|x\|\leq R, 0\leq t\leq \mathtt{T}} \left\| \frac{\mathbb{E}[X_{t+\eta}^{(\eta)} - X_t^{(\eta)} | X_t^{(\eta)} = x]}{\eta} - b(x,t) \right\| = 0 \tag{3.2}$$

and

$$\lim_{\eta\to 0} \sup_{\|x\|\leq R, 0\leq t\leq \mathtt{T}} \left\| \frac{\mathbb{E}[(X_{t+\eta}^{(\eta)} - X_t^{(\eta)})(X_{t+\eta}^{(\eta)} - X_t^{(\eta)})^T | X_t^{(\eta)} = x]}{\eta} - a(x,t) \right\| = 0. \tag{3.3}$$

Moreover, there exists a positive δ such that for all $j = 1, ..., d$ we have

$$\lim_{\eta\to 0} \sup_{\|x\|\leq R, 0\leq t\leq \mathtt{T}} \frac{\mathbb{E}[|\langle X_{t+\eta}^{(\eta)} - X_t^{(\eta)}, e_j\rangle|^{2+\delta} | X_t^{(\eta)} = x]}{\eta} = 0; \tag{3.4}$$

here, $\{e_1, ..., e_d\}$ denotes the canonical basis of $\mathbb{R}^d$.

Assumption 3.2. *There exists a continuous mapping $\sigma : \mathbb{R}^d \times [0,\infty[\to S_d$ such that for all $x \in \mathbb{R}^d$ and $t \geq 0$ we have $a(x,t) = \sigma(x,t)\sigma(x,t)^T$.*

Assumption 3.3. *Weak uniqueness holds for the Itô SDE*

$$\begin{cases} dX_t = b(X_t, t)dt + \sigma(X_t, t)dB_t, \quad t > 0; \\ X_0 = x_0. \end{cases} \tag{3.5}$$

Here, $\{B_t\}_{t\geq 0}$ denotes a standard d-dimensional Brownian motion.

We are now ready to state the following theorem taken from [Nelson, 1990].

Theorem 3.4. *Under Assumptions 3.1-3.3 the continuous time stochastic process $\{X_t^{(\eta)}\}_{t\geq 0}$ converges in distribution, as η tends to zero, to the unique weak solution $\{X_t\}_{t\geq 0}$ of the Itô SDE* (3.5).

3.2 Proof of the main result

To apply Theorem 3.4 we have to verify the validity of Assumptions 3.1-3.3 for $\{X_t^{(\eta)}\}_{t\geq 0}$ being defined by (3.1) with $\{x_{\eta k}^{(\eta)}\}_{k\geq 0}$ from (2.5).

- *Verification of Assumption 3.1*: fix $t > 0$ and assume that $k\eta \leq t < (k+1)\eta$ for some $k \in \mathbb{N}$. Then, according to (3.1) and (2.5) we can write

$$\begin{aligned} X_{t+\eta}^{(\eta)} - X_t^{(\eta)} &= x_{(k+1)\eta}^{(\eta)} - x_{k\eta}^{(\eta)} \\ &= -\nabla f(x_{\eta k}^{(\eta)})\eta + \overline{V_{\eta k}^{(\eta)}}\sqrt{\eta} \end{aligned}$$

and hence

$$\begin{aligned}\frac{\mathbb{E}[X_{t+\eta}^{(\eta)} - X_t^{(\eta)} | X_t^{(\eta)} = x]}{\eta} &= \frac{\mathbb{E}[-\nabla f(x_{\eta k}^{(\eta)})\eta + \overline{V_{\eta k}^{(\eta)}}\sqrt{\eta} | x_{\eta k}^{(\eta)} = x]}{\eta} \\ &= \mathbb{E}[-\nabla f(x_{\eta k}^{(\eta)}) | X_t^{(\eta)} = x] + \frac{\mathbb{E}[\overline{V_{\eta k}^{(\eta)}} | x_{\eta k}^{(\eta)} = x]}{\sqrt{\eta}} \\ &= \mathbb{E}[-\nabla f(x_{\eta k}^{(\eta)}) | x_{\eta k}^{(\eta)} = x] \\ &= -\nabla f(x).\end{aligned}$$

Therefore, if we set $b(t,x) := -\nabla f(x)$ condition (3.2) is trivially satisfied. Moreover, a similar computation gives

$$\begin{aligned}&\frac{\mathbb{E}[(X_{t+\eta}^{(\eta)} - X_t^{(\eta)})(X_{t+\eta}^{(\eta)} - X_t^{(\eta)})^T | X_t^{(\eta)} = x]}{\eta} \\ &= \frac{\mathbb{E}[(-\nabla f(x_{\eta k}^{(\eta)})\eta + \overline{V_{\eta k}^{(\eta)}}\sqrt{\eta})(-\nabla f(x_{\eta k}^{(\eta)})\eta + \overline{V_{\eta k}^{(\eta)}}\sqrt{\eta})^T | x_{\eta k}^{(\eta)} = x]}{\eta} \\ &= \nabla f(x)(\nabla f(x))^T \eta + \mathbb{E}[\overline{V_{\eta k}^{(\eta)}}(\overline{V_{\eta k}^{(\eta)}})^T] \\ &= \nabla f(x)(\nabla f(x))^T \eta + \overline{\eta}\Sigma(x).\end{aligned}$$

Hence, setting $a(x,t) := \overline{\eta}\Sigma(x)$ we satisfy assumption (3.3) by recalling that the assumption on the local Lipschitzianity of the gradient implies its boundedness on compact sets. We now verify (3.4): let $j \in \{1, ..., d\}$ and $\delta > 0$; then,

$$\begin{aligned}&\frac{\mathbb{E}[|\langle X_{t+\eta}^{(\eta)} - X_t^{(\eta)}, e_j\rangle|^{2+\delta} | X_t^{(\eta)} = x]}{\eta} \\ &= \frac{\mathbb{E}[|-\partial_{x_j} f(x_{\eta k}^{(\eta)})\eta + (\partial_{x_j} f(x_{\eta k}^{(\eta)}) - \partial_{x_j} f_{\gamma_{k+1}}(x_{\eta k}^{(\eta)}))\sqrt{\overline{\eta}}\sqrt{\eta}|^{2+\delta} | x_{\eta k}^{(\eta)} = x]}{\eta} \\ &\leq 2^{1+\delta}\frac{\mathbb{E}[|-\partial_{x_j} f(x_{\eta k}^{(\eta)})\eta|^{2+\delta} + |(\partial_{x_j} f(x_{\eta k}^{(\eta)}) - \partial_{x_j} f_{\gamma_{k+1}}(x_{\eta k}^{(\eta)}))\sqrt{\overline{\eta}}\sqrt{\eta}|^{2+\delta} | x_{\eta k}^{(\eta)} = x]}{\eta} \\ &\leq c_{\delta,\overline{\eta}}\left(|\partial_{x_j} f(x)|^{2+\delta}\eta^{1+\delta} + \mathbb{E}[|(\partial_{x_j} f(x_{\eta k}^{(\eta)}) - \partial_{x_j} f_{\gamma_{k+1}}(x_{\eta k}^{(\eta)}))|^{2+\delta} | x_{\eta k}^{(\eta)} = x]\eta^{\frac{\delta}{2}}\right) \\ &\leq c_{\delta,\overline{\eta}}\left(|\partial_{x_j} f(x)|^{2+\delta}\eta^{1+\delta} + 2^{1+\delta}|\partial_{x_j} f(x)|^{2+\delta}\eta^{\frac{\delta}{2}} + 2^{1+\delta}\mathbb{E}[|\partial_{x_j} f_{\gamma_{k+1}}(x_{\eta k}^{(\eta)}))|^{2+\delta} | x_{\eta k}^{(\eta)} = x]\eta^{\frac{\delta}{2}}\right) \\ &\leq c_{\delta,\overline{\eta}}\left(|\partial_{x_j} f(x)|^{2+\delta}\eta^{1+\delta} + 2^{1+\delta}|\partial_{x_j} f(x)|^{2+\delta}\eta^{\frac{\delta}{2}} + 2^{1+\delta}\eta^{\frac{\delta}{2}}\frac{1}{n}\sum_{i=1}^{n}|\partial_{x_j} f_i(x)|^{2+\delta}\right).\end{aligned}$$

Here, in the first and third inequalities, we utilized the basic estimate $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ while $c_{\delta,\overline{\eta}}$ stands for a positive constant whose value may vary from line to line. Since the terms containing the partial derivatives of f are locally Lipschitz, by assumption, and thus bounded on the compact sets $\|x\| \leq R$ we conclude that passing to the limit as η tends to zero the last term above will converge to zero, thus making condition (3.4) hold true.

- *Verification of Assumption 3.2*: we know from before that $a(x,t) = \overline{\eta}\Sigma(x)$; being a covariance matrix, $\Sigma(x)$ is symmetric and positive semi-definite thus the existence of a unique symmetric and positive semi-definite matrix $\sigma(x)$ such that $a(x,t) = \overline{\eta}\sigma(x)\sigma(x)^T$ is entailed. In addition, according to Theorem 5.2.3 in [Stroock and Varadhan, 1997] the bound in (2.7) implies the local Lipschitz continuity of $\sigma(x)$.

- *Verification of Assumption 3.3*: since $b(t,x) = -\nabla f(x)$ and $\sigma(x,t) = (\overline{\eta}\Sigma(x))^{\frac{1}{2}}$, the assumption on the gradients ∇f_i together with (2.7) yield the local Lipschitz continuity for the coefficients of the Itô SDE (2.6); this entails strong (and hence weak) uniqueness.

An illustration of Theorem 2.1 can be seen in Figure 1.

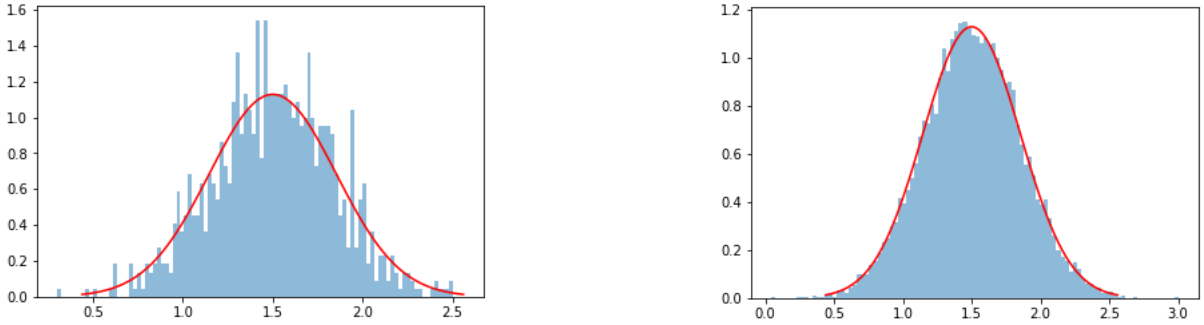

Figure 1: We repeatedly simulate the discrete time stochastic process (2.1) with $f_1 = (x-1)^2/2$ and $f_2 = (x-2)^2/2$, and hence $f(x) = (x-1)^2/4 + (x-2)^2/4$. We compare the histograms of its final value after 1000 iterations to the stationary distribution of the corresponding diffusion limit $dX_t = (1.5 - X_t)dt + (\bar{\eta})^{1/2}\frac{1}{2}dB_t$ (a mean-reverting Ornstein–Uhlenbeck process). In red the stationary distribution of the Ornstein–Uhlenbeck process (a Gaussian with mean 1.5 and variance 1/8) while in blue the histogram of the final values of the corresponding discrete time process simulated respectively 1000 (left figure) and 10000 times (right figure).

References

[Bottou, 1998] Bottou, L. (1998). On-line learning and stochastic approximations. In *In On-line Learning in Neural Networks*, pages 9–42. Cambridge University Press.

[Bottou et al., 2018] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

[Buccheri et al., 2021] Buccheri, G., Corsi, F., Flandoli, F., and Livieri, G. (2021). The continuous-time limit of score-driven volatility models. *Journal of Econometrics*, 221(2):655–675.

[Chaudhari and Soatto, 2018] Chaudhari, P. and Soatto, S. (2018). Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*.

[Corradi, 2000] Corradi, V. (2000). Reconsidering the continuous time limit of the garch(1,1) process. *Journal of Econometrics*, 96:145–153.

[Duchi et al., 2011] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

[Gower et al., 2019] Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General Analysis and Improved Rates. In *International Conference on Machine Learning*, Los Angeles, United States.

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

[Kloeden and Platen, 2011] Kloeden, P. and Platen, E. (2011). *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg.

[Li et al., 2017] Li, Q., Tai, C., and E, W. (2017). Stochastic modified equations and adaptive stochastic gradient algorithms. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2101–2110. PMLR.

[Li et al., 2019] Li, Q., Tai, C., and E, W. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47.

[Li et al., 2021] Li, Z., Malladi, S., and Arora, S. (2021). On the validity of modeling SGD with stochastic differential equations (SDEs). In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

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

[Moulines and Bach, 2011] Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

[Nelson, 1990] Nelson, D. B. (1990). Arch models as diffusion approximations. *Journal of Econometrics*, 45(1):7–38.

[Nemirovski et al., 2009] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *Society for Industrial and Applied Mathematics*, 19:1574–1609.

[Netrapalli, 2019] Netrapalli, P. (2019). Stochastic gradient descent and its variants in machine learning. *Journal of the Indian Institute of Science*, 99.

[Nguyen et al.,] Nguyen, L. M., Nguyen, P. H., Dijk, M. v., Richtárik, P., Scheinberg, K., and Takáč, M. Sgd and hogwild! convergence without the bounded gradients assumption. *Proceedings of Machine Learning Research*, 80.

[Nguyen et al., 2019] Nguyen, T. H., Simsekli, U., Gürbüzbalaban, M., and Richard, G. (2019). First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *NeurIPS*.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[Simsekli et al., 2019] Simsekli, U., Sagun, L., and Gürbüzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. *CoRR*, abs/1901.06053.

[Stroock and Varadhan, 1997] Stroock, D. and Varadhan, S. (1997). *Multidimensional Diffusion Processes*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

[Toulis and Airoldi, 2017] Toulis, P. and Airoldi, E. M. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727.

[Tsypkin and Nikolic, 1971] Tsypkin, Y. Z. and Nikolic, Z. J. (1971). *Adaptation and learning in automatic systems*, volume 73. Academic Press New York.

[Wang and Wu, 2020] Wang, Y. and Wu, S. (2020). Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *J. Mach. Learn. Res.*, 21:Paper No. 199, 103.