# Towards diffusion approximations for stochastic gradient descent without replacement

Stefan Ankirchner, Stefan Perko

# Towards diffusion approximations for stochastic gradient descent without replacement

Stefan Ankirchner* Stefan Perko†

January 16, 2022

**Abstract**

Stochastic gradient descent without replacement or reshuffling (SGDo) is predominantly used to train machine learning models in practice. However, the mathematical theory of this algorithm remains underexplored compared to its "with replacement" and "infinite data" counterparts. We propose a stochastic, continuous-time approximation to SGDo based on a family of stochastic differential equations driven by a stochastic process we call *epoched Brownian motion*, which encapsulates the behavior of reusing the same sequence of data points in subsequent epochs. We investigate this diffusion approximation by considering an application of SGDo to linear regression. Explicit convergence results are derived for constant learning rates and a sequence of learning rates satisfying the Robbins-Monro conditions. Finally, the validity of continuous-time dynamics are further substantiated by numerical experiments.

**Keywords.** Stochastic gradient descent; diffusion; stochastic differential equation; diffusion approximation; learning rate schedules; epoched Brownian motion.

*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

# 1 Introduction

Consider i.i.d. data $(x_0, y_0), \dots, (x_{N-1}, y_{N-1}) \in \mathbb{R}^d \times \mathbb{R}$ drawn from a (generalized) linear model

$$y = g(\theta^T x) + \varepsilon,$$

where $\theta \in \mathbb{R}^d$ is an unknown population parameter, $g : \mathbb{R} \to \mathbb{R}$ a function and $\varepsilon$ centered noise, independent of $x$, with finite variance $\sigma_\varepsilon^2$. We can estimate the parameter $\theta$ by performing stochastic gradient descent iterations using a squared error function $f_{(x,y)}(p) = \frac{1}{2}(g(p^T x) - y)^T (g(p^T x) - y)$ at every step. Here we assume that the data is already centered, so there is no need to fit an intercept as well.

After $N$ steps we have seen all our data and then decide to restart the next epoch with the same sequence of data points in the same order.

The dynamics of this *SGDo* process, i.e. stochastic gradient descent without replacement (or reshuffling), are then given by

$$\chi_{n+1} = \chi_n - \eta_n \nabla f_{(x_{n \bmod N}, y_{n \bmod N})}(\chi_n), \tag{1.1}$$

where $(\eta_n)_{n \in \mathbb{N}}$ is a non-increasing sequence of learning rates (LR).

Dynamics (1.1) should be contrasted with those of SGD *with replacement* given by

$$\chi_{n+1} = \chi_n - \eta_n \nabla f_{(x_{\gamma(n)}, y_{\gamma(n)})}(\chi_n), \tag{1.2}$$

where $\gamma(0), \gamma(1), \dots$ are independent and uniformly distributed on $\{0, \dots, N-1\}$, and further contrasted with SGD with *infinite data*

$$\chi_{n+1} = \chi_n - \eta_n \nabla f_{(x_n, y_n)}(\chi_n), \tag{1.3}$$

where $(x_n, y_n)_{n \in \mathbb{N}}$ is now an infinite sequence of i.i.d. data.

For the sake of this discussion we focus on online learning, but an extension to mini-batch methods is possible. Furthermore, we focus on simple linear models, i.e. we assume that $g$ is the identity function, but we believe that the following discussion gives insight beyond the linear setting.

Under the Robbins-Monro conditions

$$\sum_{n=0}^{\infty} \eta_n = \infty, \quad \sum_{n=0}^{\infty} \eta_n^2 < \infty, \tag{1.4}$$

and boundedness conditions of the data $x, y$ one can show the convergence of infinite-data SGD (1.3) to the population parameter $\theta$, say in $L_2$, while with-replacement SGD (1.2) is converges in $L_2$ to the ordinary-least squares (OLS) estimator $\hat{\theta}$ of $\theta$ (cf. [8]).

A natural question to ask is then whether convergence still holds in the SGDo setting of (1.1). In this article we show for the sequence of learning rates

$$\eta_n = \frac{1}{1+n}, n \geq 0,$$

which satisfies (1.4), that a continuous-time approximation to SGDo converges to a limiting distribution with mean and variance coinciding with the mean and variance of the OLS estimator $\hat{\theta}$. A more formal result is given in our main Theorem 2.1 (b) and experimentally substantiated in Section 5.2. Additionally, we consider the convergence properties for constant learning rates in Theorem 2.1 (a).

In order to better understand the dynamics of both (1.2) and (1.3) several authors have proposed approximating them by the dynamics of diffusion, i.e. a process satisfying a stochastic differential equation. In particular in the case of a constant learning rate $h \in (0, 1)$, Mandt et. al propose in [5] the following family of stochastic differential equations as an approximation of (1.3)

$$dX_t^h = -\nabla \mathbb{E} f_{(x_0,y_0)}(X_t^h)\, dt + \sqrt{h\Sigma}\, dW_t,$$

where $\Sigma$ is a covariance matrix based on the SGD increments and $W$ is a Brownian motion. Other works ([3], [4], [1]) further investigate the case of a non-constant LR schedule $u$ with $\eta_n = hu_{nh}$ and supply arguments for the following non-homogeneous dynamics

$$dX_t^h = -u_t \nabla \mathbb{E} f_{(x_0,y_0)}(X_t^h)\, dt + u_t \sqrt{h\Sigma}\, dW_t. \tag{1.5}$$

In these works in general the authors consider $\Sigma$ as a function of $X$. For simplicity we forgo doing this here and focus on the additive noise case in the spirit of Mandt et. al.

In the case of linear regression, $\nabla f_{\gamma(n)}$ is a linear function for all $n \in \mathbb{N}_0$ and the dynamics (1.5) simplify to

$$dX_t = -u_t \kappa (X_t - \theta)\, dt + u_t \sqrt{h}\sigma\, dW_t, \tag{1.6}$$

where $\theta \in \mathbb{R}^d$ is the *mean-reversion level*, which corresponds to the global minimum of the mean squared error function, and $\kappa, \sigma \in \mathbb{R}^{d\times d}$ are symmetric matrices, which are assumed to be positive-definite.

Analogously, we model the dynamics of (1.1) in the continuous setting with an approximation similar to (1.6). To this end, based on the given Brownian motion $W$ and a (continuous-time) epoch $T > 0$, we define

$$\hat{W}_t^T := W_{t-\lfloor t/T \rfloor T} + \lfloor t/T \rfloor\, W_T. \tag{1.7}$$

Note that on $\hat{W}^T$ is a Brownian motion on $[0, T)$ and satisfies

$$\hat{W}^T_{t+mT} = \hat{W}^T_t + mW_T, \qquad t \geq 0, n \in \mathbb{N}_0. \tag{1.8}$$

Notice that the increments of $\hat{W}$ on $[mT, (m+1)T]$ coincide with the increments of $W$ on $[0, T]$.

We call $\hat{W}^T$ a *$T$-epoched Brownian motion*. Replacing the driving path of the diffusion in (1.6) by epoched Brownian motion we arrive at the following linear stochastic differential equation with additive noise

$$dX_t = -u_t\kappa(X_t - \theta)\,dt + u_t\sqrt{h}\sigma\,d\hat{W}^T_t. \tag{1.9}$$

The driving noise $\hat{W}^T$ repeats itself when time $T$ elapses, up to a shift which makes it continuous.

Intuitively the dynamics (1.9) mirror SGDo. In the first epoch we draw data i.i.d. to estimate the direction of steepest descent. So up until then it is reasonable to use a Brownian motion to approximate these discrete dynamics in a continuous, stochastic setting as evidenced by a functional version of the central limit theorem. In the following epochs we do not draw any new data, nor do we permute the order of the data points. This is captured, in the continuous setting, by the property (1.8). Note that in (1.8) the $mW_T$-term is present to ensure that epoched Brownian motion is a continuous process.

Approximating SGD with replacement (1.2) yields an SDE of the form

$$dX_t = -u_t\mathbb{E}[\nabla f_{x_{\gamma(0)}, y_{\gamma(0)}}|x, y](X_t)\,dt + \sqrt{h}\Sigma(X_t)\,dW_t, \tag{1.10}$$

where in contrast to (1.5) the coefficients are now *random*, though independent of $W$, as they are defined in terms of the random variables $(x_0, y_0), \ldots, (x_{N-1}, y_{N-1})$. The mean and covariance matrix $\Sigma$ of the gradient noise in (1.2) is instead taken with respect to sequence of random integers $(\gamma(n))_{n\in\mathbb{N}}$, conditional on the data $x, y$. We call this kind of diffusion approximation with random coefficients depending on the data $x, y$ an *empirical* diffusion approximation of SGD. An advantage of this empirical approximation is that it requires no knowledge of the true model generating the data.

In the case of linear regression (1.10) simplifies to

$$dX_t = -u_t\hat{\kappa}(X_t - \hat{\theta})\,dt + u_t\sqrt{h}\hat{\sigma}\,dW_t, \tag{1.11}$$

where $\hat{\kappa}, \hat{\theta}$ and $\hat{\sigma}$ are functions of the data $x, y$.

An analogous empirical diffusion approximation seems to be infeasible for SGDo. To recover the without-replacement behavior on the given data

set $(x_k, y_k)_{k=0}^{N-1}$ the draws are no longer independent, nor are they identically distributed, which makes the choice of a reasonable, data-dependent diffusion driver highly non-obvious.

Previous works on SGDo have mainly focused on comparing the convergence rates of SGD with replacement (1.2) and SGDo (1.1), where empirically the latter is known to converge faster. In [7] Shamir establishes convergence results for set of algorithms enjoying regret bounds, which includes SGDo. In [6] Nagraj et. al use the method of exchangeable pairs to derive non-asymptotic convergence results for general smooth, strongly convex functions.

In general SGDo is rarely studied in the mathematical literature compared to SGD with replacement or with infinite data. However, this is likely because the latter two are mathematically easier to discuss because of i.i.d. gradient noise. On the flip side SGDo is almost universally used in machine learning practice and therefore of great importance.

# 2 Main result

To motivate and state the main results, let us first discuss more in detail how to derive equation (1.9) and how to choose the coefficients in the case of linear regression.

To ease notation consider first the one-dimensional setting $d = 1$. Then for linear regression ($g(x) = x$) (1.1) simplifies to

$$\begin{aligned}\chi_{n+1} =& \chi_n - \eta_n \nabla f_{(x_n,y_n)}(\chi_n)\\ =& \chi_n - \eta_n(\chi_n x_n^2 - x_n y_n)\\ =& (1-\eta_n x_n^2)\chi_n + \eta_n x_n y_n, \end{aligned} \tag{2.1}$$

where $(x_0, y_0), \dots, (x_{N-1}, y_{N-1}) \sim \nu$ are i.i.d. and $(x_{n+mN}, y_{n+mN}) = (x_n, y_n)$ for all $m, n \in \mathbb{N}_0$. By decomposing the steps as $n+mN$ with $n \in \{0, \dots, N-1\}$ the number $m$ specifies the current epoch.

Solving recursion (2.1), cf. Lemma 3.1 below, we have

$$\chi_n = \chi_0 \prod_{k=0}^{n-1}(1-\eta_k x_k^2) + \sum_{l=0}^{n-1}\left(\prod_{k=l+1}^{n-1}(1-\eta_k x_k^2)\right)\eta_l x_l y_l.$$

For $n > N$ the products feature dependent factors. We may rewrite this equation at a completed epoch as

$$\chi_{(m+1)N} = \chi_0 \prod_{k=0}^{N-1}\prod_{j=0}^{m}(1-\eta_{k+Nj} x_k^2) + \sum_{l=0}^{N-1}\sum_{i=0}^{m}\left(\prod_{k=l+1}^{(m+1)N-1}(1-\eta_k x_k^2)\right)\eta_l x_l y_l.$$

Computing the expectation of $\chi_{n+mN}$ is cumbersome, but even more so is it to compute its variance, as the two summands are correlated.

Our idea is to pass to the continuous-time limit, since it is considerably easier to study distributional properties of the limit than of the discrete-time process (2.1).

To this end, let us rewrite (2.1) as

$$\chi_{n+1} = \chi_n - \eta_n(\mu_x^2\chi_n - \mu_{x,y}) - \eta_n M_n,$$

where

$$\mu_x^k := \mathbb{E}[x_0^k], \mu_{x,y}^{k,l} := \mathbb{E}[x_0^k y_0^l],$$

and $\mu_{x,y} := \mu_{x,y}^{1,1}$ are (marginal) moments and joint moments associated with the distribution $\nu$, and

$$M_n := (x_n^2 - \mu_x^2)\chi_n - x_n y_n + \mu_{x,y}, n \in \mathbb{N}_0$$

Then for $n \in \{0, \dots, N-1\}$ we have $\mathbb{E}[M_n|\chi_n] = 0$ and so the conditional variance matrix satisfies

$$\begin{aligned}\mathrm{Var}[M_n|\chi_n] =& \mathbb{E}[(x_n^2 - \mu_x^2)^2\chi_n^2 - 2(x_n^2 - \mu_x^2)(x_n y_n - \mu_{x,y})\chi_n + (x_n y_n - \mu_{x,y})^2|\chi_n] \\ =& (\mu_x^4 - (\mu_x^2)^2)\chi_n^2 - 2(\mu_{x,y}^{3,1} - \mu_{x,y}\mu_x^2)\chi_n + \mu_{x,y}^{2,2} - (\mu_{x,y})^2.\end{aligned}$$

For a diffusion approximation it is therefore natural to choose the state-dependent diffusion coefficient $\sigma(p) = \sqrt{\mathrm{Var}[M_n|\chi_n = p]}$. However, we want to use a constant coefficient instead and in this case there is some flexibility in choosing $\sigma$.

In general

$$\begin{aligned}\mu_{x,y}^{3,1} =& \mathbb{E}[x_0^3(\theta x_0 + \varepsilon)] = \theta\mu_x^4, \\ \mu_{x,y} =& \mathbb{E}[x_0(\theta x_0 + \varepsilon)] = \theta\mu_x^2, \\ \mu_{x,y}^{2,2} =& \mathbb{E}[x_0^2(\theta x_0 + \varepsilon)^2] = \theta^2\mu_x^4 + \mu_x^2\sigma_\varepsilon^2,\end{aligned}$$

and so

$$\sigma(p)^2 = (\mu_x^4 - (\mu_x^2)^2)(p - \theta)^2 + \mu_x^2\sigma_\varepsilon^2.$$

A decent constant approximation is given by setting $p = \theta$ so that

$$\sigma = \sqrt{\sigma_\varepsilon^2\mu_x^2}.$$

Since the drift coefficient $-\mu_x^2(p - \theta)$ pushes us to $\theta$ exponentially fast it is more important that the diffusion term is accurate around the mean-reversion level $\theta$ than anywhere else. We shall set

$$\kappa := \mu_x^2, \sigma := \sqrt{\sigma_\varepsilon^2\mu_x^2}.$$

Then we can rewrite (2.1) as

$$\chi_{n+1} = \chi_n - \eta_n \kappa(\chi_n - \theta) - \eta_n M_n, \tag{2.2}$$

where $\mathrm{Var}[M_n|\chi_n] \approx \sigma^2$ for $n \in \{0, \dots, N-1\}$ and $\chi_n$ close to $\theta$.

Analogously, for $d > 1$ by defining the symmetric and positive semi-definite matrices

$$\kappa := \mu_x^2 = \mathbb{E}[x_0 x_0^T], \quad \sigma := \sqrt{\sigma_\varepsilon^2 \mu_x^2}$$

we have

$$\chi_{n+1} = \chi_n - \eta_n \kappa(\chi_n - \theta) - \eta_n M_n,$$

where the conditional covariance matrix $\mathrm{Cov}[M_n|\chi_n]$ is close to $\sigma^2$ for $n \in \{0, \dots, N-1\}$ and $\chi_n$ close to $\theta$. Here $\sqrt{\mu_x^2}$ is the unique symmetric and positive semi-definite matrix, such that $\sqrt{\mu_x^2}\sqrt{\mu_x^2} = \mu_x^2$.

By decomposing $\eta_n = h u_{nh}^h$ for an initial learning rate $h \in (0,1)$ and a learning rate schedule $u : (0,1) \times [0,\infty) \to [0,1], (h,t) \mapsto u_t^h$, bounded in $h$ and $t$, we can see that the iterations (2.2) with $n \in \{mN, \dots, (m+1)N-1\}$ restricted to a single epoch and started at a *deterministic* value $\chi_{mN} \in \mathbb{R}^d$ at the beginning of said epoch are well approximated by the solution to the $h$-indexed family of linear stochastic differential equations

$$dX_t^h = -\kappa u_t (X_t^h - \theta)\, dt + \sqrt{h} u_t \sigma\, dW_t,$$

with $t \in [mT, (m+1)T)$, $X_{mT} = \chi_{mN}$, $W$ is a $d$-dimensional Brownian motion and $T = Nh$ (cf. e.g. [5]).

By extension we expect that the iterations (2.2) are well approximated by the solution to the sequence of stochastic differential equations

$$dX_{t+mT} = -u_{t+mT}\kappa(X_{t+mT} - \theta)\, dt + \sqrt{h} u_{t+mT}\sigma\, dW_t, t \in [0,T). \tag{2.3}$$

More succinctly we may describe $X$ as the solution to the equation 1.9.

We now state our main result. To this end we recall that for any symmetric and positive-definite matrix $A$ the hyperbolic cotangent of $A$ is defined, and given by

$$\coth A := (e^{2A} - 1)^{-1}(e^{2A} + 1).$$

**Theorem 2.1.** *Let $X$ be the solution to the linear stochastic differential equation*

$$dX_t = -u_t \mu_x^2 (X_t - \theta)\, dt + u_t \sqrt{h \sigma_\varepsilon^2 \mu_x^2}\, d\hat{W}_t^{Nh} \tag{2.4}$$

*driven by $Nh$-epoched Brownian motion. Then we have the following.*

*(a) If $u = 1$ (a constant), then $X$ converges in distribution*

$$X_t \to \mathcal{N}\left(\theta, \frac{1}{2}\sigma_\varepsilon^2 h\sqrt{\mu_x^2}\coth(\mu_x^2 Nh/2)(\mu_x^2)^{-1}\sqrt{\mu_x^2}\right),$$

*as $t \to \infty$.*

*(b) If $u_t = \frac{1}{1+\frac{t}{h}}$, so that $u_{nh} = \frac{1}{1+n}$, and $h^{-1}$ is not an eigenvalue of $\mu_x^2$, then $X$ converges in distribution*

$$X_t \to \mathcal{N}\left(\theta, \frac{\sigma_\varepsilon^2}{N}(\mu_x^2)^{-1}\right),$$

*as $t \to \infty$.*

Some remarks on Theorem 2.1 are in order.

According to the Taylor approximation

$$\coth x = x^{-1} + \frac{x}{3} + \mathcal{O}(x^3)$$

the variance of the limiting distribution in (a) is

$$\begin{aligned}&\frac{1}{2}\sigma_\varepsilon^2 h\sqrt{\mu_x^2}\left(\frac{2}{Nh}(\mu_x^2)^{-1} + \frac{Nh\mu_x^2}{2} + \mathcal{O}(N^3h^3)\right)(\mu_x^2)^{-1}\sqrt{\mu_x^2}\\ =&\frac{\sigma_\varepsilon^2}{N}(\mu_x^2)^{-1} + \mathcal{O}(Nh^2)\end{aligned}$$

as $h \downarrow 0$. This should be compared with the fact the OLS estimator[1]

$$\hat{\theta} = \frac{\sum_{n=1}^N x_n y_n}{\sum_{n=1}^N x_n^2}$$

has mean $\theta$ and variance $\frac{\sigma_\varepsilon^2}{N}(\mu_x^2)^{-1}$. Thus, for small learning rates $X_t$ attains the variance of the minimum-variance unbiased estimator (MVUE) in the limit $t \to \infty$.

Similarly, the limiting distribution for $u_t^h = \frac{1}{1+\frac{t}{h}}$ in (b) has the same mean and variance as the OLS estimator. Since the OLS estimator is the MVUE one may be tempted to conclude that in fact $X_t \to \hat{\theta}$ in (b), as $t \to \infty$ in some sense, say almost surely. This cannot technically be true, since $X_t$ is not a function of the data $(x_n, y_n)_{n=0}^{N-1}$ and therefore not a point estimator at all. Nevertheless, in spirit this is what Theorem 2.1 (b) tells us.

[1] $\hat{\theta}$ is the OLS estimator according to the Gauss-Markov theorem, as long as we use our assumption that $x$ and $y$ are centered.

# 3 Linear differential equations driven by epoched Brownian motion

Let $(\Omega, \mathcal{F}_\infty, \mathbb{P})$ be a complete probability space, $d \in \mathbb{N}$ and $W$ be a $d$-dimensional Brownian motion defined on $\Omega$. Consider again the definition of $T$-epoched Brownian motion in (1.7) and the proposed diffusion approximation of (1.1) given by

$$dX_t = -u_t\kappa(X_t - \theta)\,dt + u_t\sigma\,d\hat{W}_t^T, \tag{3.1}$$

where $\kappa, \sigma \in \mathbb{R}^{d\times d}$ are symmetric and positive-definite, and $\theta \in \mathbb{R}^d$.

For ease of notation we have absorbed the $\sqrt{h}$-term into $\sigma$. Let $m \in \mathbb{N}_0$. Then in epoch $m$ we can write equation (3.1) as the linear stochastic differential equation

$$dX_{t+mT} = -u_{t+mT}\kappa(X_{t+mT} - \theta)\,dt + u_{t+mT}\sigma\,dW_t, t \in [0, T). \tag{3.2}$$

driven by the Brownian motion $W$. If the initial condition $X_{mT}$ was deterministic, then the solution of (3.2) is known to be given by

$$\begin{aligned} X_{t+mT} = f_{t+mT}f_{mT}^{-1}X_{mT} &+ \left(\int_0^t f_{t+mT}f_{s+mT}^{-1}u_{s+mT}\,ds\right)\kappa\theta \\ &+ \left(\int_0^t f_{t+mT}f_{s+mT}^{-1}u_{s+mT}\,dW_s\right)\sigma, \end{aligned} \tag{3.3}$$

where $f_t = \exp(-\kappa\int_0^t u_s\,ds)$ and exp denotes the matrix exponential. Observe that the RHS of (3.3) is meaningful also for arbitrary non-deterministic $X_{mT}$. Therefore, we use the pathwise representation (3.3) for defining a solution of (3.1). Notice that for $m \geq 1$ the initial value $X_{mT}$ of (3.2) is not independent of the driving Brownian motion. Thus, the process (3.3) is not a solution of (3.2) in the Itô sense with respect to $\mathcal{F}$, the augmented filtration generated by $W$. We remark, however, that (3.3) can be be shown to be a solution in the Itô sense under the *enlarged* filtration $(\mathcal{F}_t \vee \sigma(X_{mT}))_{t\in[0,T)}$.

At every completed epoch $m \in \mathbb{N}_0$ we have

$$\begin{aligned} X_{(m+1)T} = f_{(m+1)T}f_{mT}^{-1}X_{mT} &+ \left(\int_0^T f_{t+mT}f_{s+mT}^{-1}u_{s+mT}\,ds\right)\kappa\theta \\ &+ \left(\int_0^T f_{t+mT}f_{s+mT}^{-1}u_{s+mT}\,dW_s\right)\sigma. \end{aligned} \tag{3.4}$$

The following lemma is standard and well-known, but repeated here for the reader's convenience.

**Lemma 3.1.** *Let $A : \mathbb{N}_0 \to \mathbb{R}^{d\times d}$ and $x, B : \mathbb{N}_0 \to \mathbb{R}^d$ be sequences. Then the linear recurrence*

$$x_{n+1} = A_n x_n + B_n$$

*has the solution*

$$x_n = x_0 \prod_{k=0}^{n-1} A_k + \sum_{l=0}^{n-1} \left( \prod_{k=l+1}^{n-1} A_k \right) B_l.$$

*Proof.* The statement holds for $n = 0$. Suppose it holds for $n \in \mathbb{N}$. Then,

$$\begin{aligned} x_{n+1} =& A_n x_n + B_n \\ =& x_0 \prod_{k=0}^{n} A_k + \sum_{l=0}^{n-1} \left( \prod_{k=l+1}^{n} A_k \right) B_l + B_n \\ =& x_0 \prod_{k=0}^{n} A_k + \sum_{l=0}^{n} \left( \prod_{k=l+1}^{n} A_k \right) B_l. \end{aligned}$$

□

We next apply Lemma 3.1 to equation (3.4). First note that we have

$$\prod_{k=l}^{m-1} f_{(k+1)T} f_{kT}^{-1} = f_{mT} f_{lT}^{-1}.$$

We define

$$\begin{aligned} U_T^m &:= \sum_{l=0}^{m-1} f_{mT} f_{(l+1)T}^{-1} \int_0^T f_{(l+1)T} f_{s+lT}^{-1} u_{s+lT}\, ds = \sum_{l=0}^{m-1} \int_0^T u_{s+lT} f_{mT} f_{s+lT}^{-1}\, ds, \\ V_T^m &:= \sum_{l=0}^{m-1} f_{mT} f_{(l+1)T}^{-1} \int_0^T f_{(l+1)T} f_{s+lT}^{-1} u_{s+lT}\, dW_s = \sum_{l=0}^{m-1} \int_0^T u_{s+lT} f_{mT} f_{s+lT}^{-1}\, dW_s. \end{aligned}$$

By applying Lemma 3.1 we arrive at the following explicit epoch-wise dynamics.

**Lemma 3.2.** *The solution $X$ of (3.1) at the end of epoch $m \in \mathbb{N}_0$ is given by*

$$X_{mT} = f_{mT} X_0 + U_T^m \kappa\theta + V_T^m \sigma. \tag{3.5}$$

The lemma implies that the solution $X$ is a Gaussian process with

$$\mathbb{E} X_{mT} = f_{mT} X_0 + U_T^m \kappa\theta,$$

and

$$\begin{aligned}\operatorname{Var} X_{mT} =& \sigma \operatorname{Var} V_T^n \sigma \\ =& \sigma \left( \int_0^T \left( \sum_{k=0}^{m-1} u_{s+kT} f_{mT} f_{s+kT}^{-1} \right) \left( \sum_{l=0}^{m-1} u_{s+lT} f_{mT} f_{s+lT}^{-1} \right)^T ds \right) \sigma \\ =& \sigma \left( \int_0^T \sum_{k,l=0}^{m-1} u_{s+kT} u_{s+lT} f_{mT} f_{s+kT}^{-1} f_{s+lT}^{-T} f_{mT}^T \, ds \right) \sigma.\end{aligned}$$

# 4 Proof of the main result

By Lemma 3.2 the solution to (3.1) at the end of any completed epoch is given by

$$X_{mT} = f_{mT} X_0 + U_T^m \kappa\theta + V_T^m \sigma, \tag{4.1}$$

where

$$U_T^m = \sum_{l=0}^{m-1} \int_0^T u_{s+lT} f_{mT} f_{s+lT}^{-1} \, ds, \quad V_T^m = \sum_{l=0}^{m-1} \int_0^T u_{s+lT} f_{mT} f_{s+lT}^{-1} \, dW_s,$$

and

$$f_t = \exp\left( -\kappa \int_0^t u_s \, ds \right).$$

If $\int_0^\infty u_s \, ds = \infty$ and $\kappa$ is positive-definite, then $f_{mT} \to 0$, as $m \to \infty$. The condition $\int_0^\infty u_s \, ds = \infty$ is fulfilled for $u = 1$ and $u_t = \frac{1}{1+at}$ with $a > 0$.

It remains to discuss the behavior of $U_T^m$ and $V_T^m$ as $m \to \infty$ in these cases. We will prove the following proposition. Our main result Theorem 2.1 follows by simply plugging $T = Nh$ and considering the specific coefficients in (2.4), whereas Proposition 4.1 is meaningful regardless of our statistical motivation.

**Proposition 4.1.** *Let $X$ be the solution to* (3.1)*. Then we have the following.*

*(a) If $u = 1$ (a constant), then $X$ converges in distribution*

$$X_t \to \mathcal{N}\left( \theta, \frac{1}{2} \sigma \coth(\kappa T/2) \kappa^{-1} \sigma \right),$$

*as $t \to \infty$.*

*(b) If $u_t = \frac{1}{1+at}$ and $a$ is not an eigenvalue of $\kappa$, then $X$ converges in distribution*

$$X_t \to \mathcal{N}\left( \theta, \frac{1}{T} \sigma \kappa^{-2} \sigma \right),$$

*as $t \to \infty$.*

## 4.1 Proof of main result

In the case $u = 1$, we have

$$U_T^m = \sum_{l=0}^{m-1} \int_0^T f_{mT} f_{s+lT}^{-1} \, ds, \quad V_T^m = \sum_{l=0}^{m-1} \int_0^T f_{mT} f_{s+lT}^{-1} \, dW_s,$$

and $f_t = \exp(-\kappa t)$. Then,

$$\begin{aligned} U_T^m &= \sum_{l=0}^{m-1} \int_0^T e^{-\kappa(mT-(s+lT))} \, ds \\ &= \sum_{l=0}^{m-1} (e^{\kappa T} - 1) e^{-(m-l)\kappa T} \kappa^{-1} \\ &= (1 - e^{-\kappa m T}) \kappa^{-1} \\ &\longrightarrow \kappa^{-1}, \end{aligned}$$

as $m \to \infty$. Similarly,

$$\begin{aligned} \operatorname{Var} V_T^m =& \sum_{k,l}^{m-1} \int_0^T e^{-\kappa(mT-(s+kT))} e^{-\kappa(mT-(s+lT))} \, ds \\ =& \frac{1}{2} \sum_{k,l}^{m-1} e^{-\kappa(2m-k-l)T} (e^{2\kappa T} - 1) \kappa^{-1} \\ =& \frac{1}{2} e^{-2\kappa m T} \left( \sum_{l=0}^{m-1} e^{\kappa l T} \right)^2 (e^{2\kappa T} - 1) \kappa^{-1} \\ =& \frac{1}{2} (e^{\kappa T} - 1)^{-2} e^{-2\kappa m T} (e^{\kappa m T} - 1)^2 (e^{2\kappa T} - 1) \kappa^{-1} \\ =& \frac{1}{2} (e^{\kappa T} - 1)^{-2} (1 - 2e^{-\kappa m T} + e^{-2\kappa m T}) (e^{2\kappa T} - 1) \kappa^{-1}, \end{aligned}$$

where we have repeatedly used commutativity of the given matrices. Letting $m \to \infty$, we obtain

$$\begin{aligned} \lim_{m \to \infty} V_T^m =& \frac{1}{2} (e^{\kappa T} - 1)^{-2} (e^{\kappa T} - 1)(e^{\kappa T} + 1) \kappa^{-1} \\ =& \frac{1}{2} \coth(\kappa T / 2) \kappa^{-1}, \end{aligned}$$

where

$$\coth A = (e^{2A} - 1)^{-1} (e^{2A} + 1),$$

for any symmetric and positive-definite matrix $A \in \mathbb{R}^{d\times d}$. In total,

$$X_t \to \mathcal{N}\left(\theta, \frac{1}{2}\sigma \coth(\kappa T/2)\kappa^{-1}\sigma\right).$$

This proves Proposition 4.1 (a) and by extension Theorem 2.1 (a).

Now, consider (b), i.e.

$$u_t = \frac{1}{1+at}, t \geq 0$$

for some $a > 0$. By defining

$$c^A := \exp((\log c)A)$$

for every matrix $A \in \mathbb{R}^{d\times d}$ and $c > 0$, we have $\int_0^t u_s\, ds = \frac{1}{a}\log(1+at)$ and

$$f_t = (1+at)^{-\frac{\kappa}{a}}.$$

Recalling the dynamics of $X_{mT}$ in (4.1) we compute

$$\begin{aligned}\sum_{l=0}^{m-1} u_{s+lT} f_{mT} f_{s+lT}^{-1} &= \sum_{l=0}^{m-1}(1+a(s+lT))^{-1}(1+amT)^{-\frac{\kappa}{a}}(1+a(s+lT))^{\frac{\kappa}{a}} \\ &= (1+amT)^{-\frac{\kappa}{a}} \sum_{l=0}^{m-1}(1+a(s+lT))^{\frac{\kappa}{a}-1_{d\times d}}, \qquad (4.2)\end{aligned}$$

since $c^a 1_{d\times d} = c^{a1_{d\times d}}$ for all $a \in \mathbb{R}$ and $c > 0$. For ease of notation we write $A - c := A - c1_{d\times d}$ for any $A \in \mathbb{R}^{d\times d}$ and $c \in \mathbb{R}$.

We need the following well-known lemma to compute exactly the limits of $U_T^m$ and $V_T^m$ as $m \to \infty$.

**Lemma 4.2.** *Let $a \leq b \in \mathbb{N}_0$ and $f : [a,b] \to \mathbb{R}^{d\times d}$ be a matrix-valued function in $C^1$, i.e. continuously differentiable. Then,*

$$\sum_{n=a+1}^{b} f(n) = \int_a^b f(r)\, dr + \int_a^b \{r\} f'(r)\, dr,$$

*where $\{x\} = x - \lfloor x \rfloor$ is the fractional part of $x \in \mathbb{R}$.*

*Proof.* We have

$$\begin{aligned}\int_n^{n+1} \{r\} f'(r)\, dr =& \int_n^{n+1} (r-n) f'(r)\, dr \\ =& \int_n^{n+1} r f'(r)\, dr - n \int_n^{n+1} f'(r)\, dr \\ =& (n+1) f(n+1) - n f(n) - \int_n^{n+1} f(r)\, dr - n \int_n^{n+1} f'(r)\, dr \\ =& f(n+1) - \int_n^{n+1} f(r)\, dr,\end{aligned}$$

and so

$$\begin{aligned}\sum_{n=a+1}^{b} f(n) =& \sum_{n=a}^{b-1} f(n+1) \\ =& \sum_{n=a}^{b-1} \int_n^{n+1} f(r)\, dr + \sum_{n=a}^{b-1} \int_n^{n+1} \{r\} f'(r)\, dr \\ =& \int_a^b f(r)\, dr + \int_a^b \{r\} f'(r)\, dr.\end{aligned}$$

□

Clearly, the function

$$(0,\infty) \to \mathbb{R}^{d\times d}, t \to (c+at)^A = \exp(\log(c+at)A)$$

is $C^1$, with derivative given by

$$\begin{aligned}\partial_t \exp(\log(c+at)A) =& \partial_t \log(c+at) A \exp(\log(c+at)A) \\ =& \frac{a}{c+at} A (c+at)^A \\ =& aA(c+at)^{A-1},\end{aligned}$$

for all $c, a > 0$ and $A \in \mathbb{R}^{d\times d}$. Therefore, for all $n \in \mathbb{N}$

$$\begin{aligned}\sum_{l=1}^{m-1} (1+a(s+lT))^{\frac{\kappa}{a}-1} =& \int_0^{m-1} (1+a(s+rT))^{\frac{\kappa}{a}-1}\, dr \\ &+ \int_0^{m-1} \{r\} aT \left(\frac{\kappa}{a} - 1\right) (1+a(s+rT))^{\frac{\kappa}{a}-2}\, dr.\end{aligned} \tag{4.3}$$

Consider arbitrary $c_1, c_2 \in \mathbb{R}, b > 0$ and a matrix $A \in \mathbb{R}^{d\times d}$. Assuming 1 is not an eigenvalue of $A$,

$$\begin{aligned}\int_0^{m-1} (c_2+bm)^{-A}(c_1+br)^{A-2}\,dr =& \frac{1}{b}(A-1)^{-1}(c_1+br)^{A-1}\Big|_{r=0}^{r=m-1}(c_2+bm)^{-A}\\ =& \frac{(A-1)^{-1}}{bc_1+b^2(m-1)}\left(\frac{c_1+b(m-1)}{c_2+bm}\right)^A\\ &-\frac{(A-1)^{-1}}{bc_1}\left(\frac{c_1}{c_2+bm}\right)^A\\ &\longrightarrow 0, \end{aligned} \tag{4.4}$$

as $m \to \infty$. Similarly if $A$ is invertible, we have

$$\begin{aligned}\int_0^{m-1} (c_2+bm)^{-A}(c_1+br)^{A-1}\,dr =& \frac{1}{b}A^{-1}(c_1+br)^{A}\Big|_{r=0}^{r=m-1}(c_2+bm)^{-A}\\ =& \frac{A^{-1}}{b}\left(\frac{c_1+b(m-1)}{c_2+bm}\right)^A\\ &-\frac{A^{-1}}{b}\left(\frac{c_1}{c_2+bm}\right)^A\\ &\longrightarrow \frac{1}{b}A^{-1}, \end{aligned} \tag{4.5}$$

as $m \to \infty$.

Applying (4.4) and (4.5) to (4.2) and (4.3), and taking into account the assumption that 1 is not an eigenvalue of $\frac{\kappa}{a}$ and that $\kappa$ is invertible, we get

$$\begin{aligned}\lim_{m\to\infty}\sum_{l=0}^{m-1} u_{s+lT}f_{mT}f_{s+lT}^{-1} =& \lim_{m\to\infty}(1+amT)^{-\frac{\kappa}{a}}\sum_{l=1}^{m-1}(1+a(s+lT))^{\frac{\kappa}{a}-1}\\ &+\lim_{m\to\infty}(1+amT)^{-\frac{\kappa}{a}}(1+as)^{\frac{\kappa}{a}-1}\\ =& \frac{1}{aT}\left(\frac{\kappa}{a}\right)^{-1}\\ =& \frac{1}{T}\kappa^{-1}. \end{aligned} \tag{4.6}$$

Note that the Frobenius norm of the matrix exponential of a symmetric and positive-semidefinite matrix $A$ satisfies

$$\left\|e^A\right\|_2^2 = \mathrm{tr}[e^A(e^A)^T] = \mathrm{tr}[e^{2A}] = \sum_{k=1}^{d} e^{2\lambda_k} > 0,$$

where $\lambda_1, \dots, \lambda_d \geq 0$ are the eigenvalues of $A$. In particular,

$$\left\| \left( \frac{c_1 + b(m-1)}{c_2 + bm} \right)^A \right\|_2 = \sqrt{\sum_{k=1}^{d} \exp\left( 2 \log\left( \frac{c_1 + b(m-1)}{c_2 + bm} \right) \lambda_k \right)}$$

for $m \geq 2$ is uniformly bounded by its value in $m = 2$. Similarly,

$$\left\| \left( \frac{c_1}{c_2 + bm} \right)^A \right\|_2$$

is uniformly bounded by its value in $m = 1$. Hence, the integrals in (4.4) and (4.5) are uniformly bounded by integrable matrix-valued functions and analogously so is the term $(1 + amT)^{-\frac{\kappa}{a}}(1 + as)^{\frac{\kappa}{a}-1}$ in (4.6).

Therefore, we can apply dominating convergence to get

$$U_T^n \to \int_0^T \frac{1}{T} \kappa^{-1} \, dt = \kappa^{-1},$$

as $m \to \infty$. Recalling, equation (4.1)

$$X_{mT} = f_{mT} X_0 + U_T^m \kappa\theta + V_T^m \sigma$$

and that $f_{mT} \to 0$, we can conclude

$$\mathbb{E} X_{mT} \to \theta,$$

as $m \to \infty$, i.e. in expectation $X$ converges to its mean-reversion level $\theta$. Further, again using (4.6) and dominated convergence we have

$$\begin{aligned}
\operatorname{Var} V_T^m =& \operatorname{Var}\left( \sum_{l=0}^{m-1} \int_0^T u_{s+lT} f_{mT} f_{s+lT}^{-1} \, dW_s \right) \\
=& \int_0^T \left( \sum_{l=0}^{m-1} u_{s+lT} f_{mT} f_{s+lT}^{-1} \right) \left( \sum_{l=0}^{m-1} u_{s+lT} f_{s+lT}^{-T} f_{mT}^{T} \right) ds \\
\to& \int_0^T \frac{1}{T^2} \kappa^{-1} \kappa^{-T} \, ds \\
=& \frac{\kappa^{-2}}{T},
\end{aligned}$$

as $m \to \infty$. So in total

$$X_t \to \mathcal{N}\left( \theta, \frac{1}{T} \sigma^T \kappa^{-2} \sigma \right),$$

in distribution as $t \to \infty$, proving Proposition 4.1 (b) and Theorem 2.1 (b).

# 5 Experiments

By performing experiments one can find evidence that (2.4) is indeed a decent approximation to SGDo and that the results of Theorem 2.1 can be in some sense applied to SGDo as well. We will focus on the LR schedule

$$u_t = \frac{1}{1 + \frac{t}{h}}, t \geq 0.$$

## 5.1 Weak approximation error

Fix $T = 1$. For every initial learning rate $h \in (0, 1)$ we draw $N = \lfloor T/h \rfloor$ i.i.d. data points according to the linear model

$$y = -x + 0.5\varepsilon,$$

where $x, \varepsilon \sim \mathcal{N}(0, 1)$ are independent.

We estimate the quadratic deviation from the population parameter $\theta = -1$,

$$\mathbb{E}g(Y_N) := \mathbb{E}[(Y_N - (-1))^2],$$

for $Y = \chi, X, X^0$, where $\chi$ is given by (1.1), $X$ is given by (1.9) and $X^0$ is the gradient flow approximation given by the ordinary differential equation

$$\dot{X}_t^0 = -\mu_x^2 u_t (X_t^0 - \theta), t \geq 0. \tag{5.1}$$

For $Y = X^0$ we can compute the deviation exactly, where as for $Y = \chi, X$ we use Monte-Carlo sampling with $M = 500,000$ instances. Finally, we compute the weak errors

$$|\mathbb{E}g(\chi_N^h) - \mathbb{E}g(X_T^h)|, |\mathbb{E}g(\chi_N^h) - g(X_T^0)|.$$

The result of one simulation is depicted in Figure 1.

Observe that, the weak error associated with $\chi$ and $X$ is lower for *small* learning rates $h$, but can exceed the weak error for gradient flow for $h$ close to 1. Since weak approximations are only accurate for small $h$ the diffusion $X$ can still be considered to be a better weak approximation to $\chi$ than the gradient flow approximation $X^0$.

## 5.2 Limiting distribution of SGDo vs. OLS estimator

For epoch lengths $N = 2^n, n \in \{0, \ldots, 15\}$ we shall draw $N$ i.i.d. data points according to the linear model

$$y = -x + 0.5\varepsilon,$$

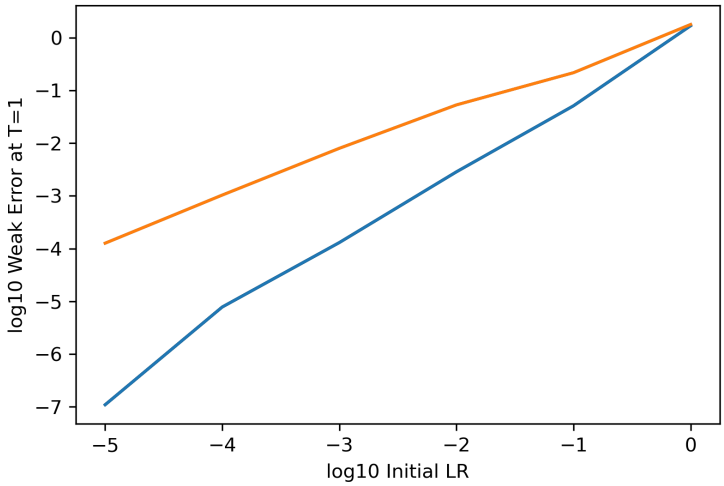


Figure 1: Weak approximation error of the diffusion approximation (2.4) (blue) vs. weak approximation error of the gradient flow approximation (5.1) (orange)

where $x, \varepsilon \sim \mathcal{N}(0,1)$ are independent. Then we run stochastic gradient descent without replacement or reshuffling with epoch length $N$ for $\max(100N, 10000)$ iterations.

In order to approximate the standard deviation of the limiting distribution, i.e. $s > 0$ with

$$\chi_n \to \mathcal{N}(\theta, s^2)$$

we estimate $s$ using $M = 2000$ Monte-Carlo instances and time points $n = (m-1)N, \dots, mN, \dots, mN + r$, where $n = mN + r$ with $n, r$ coprime. We compare the estimated standard deviations with the standard deviation of the OLS estimator for the population parameter $\theta = -1$ given by

$$N \mapsto \sqrt{\frac{\operatorname{Var}[0.5\varepsilon]}{N \operatorname{Var}(x)}} = \frac{1}{2\sqrt{N}}.$$

Here we consider the theoretical variance of $x$, rather than the sample variance of $x_0, \dots, x_{N-1}$. The result is plotted in Figure 2. It confirms that the mean and variance of the limiting distribution of $X$ in 2.1 (b) are also attained by SGDo.

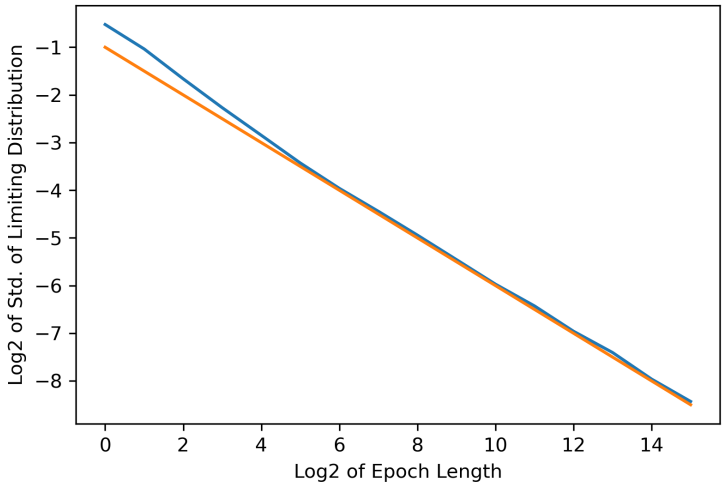


Figure 2: Standard deviation of the limiting distribution of SGDo (blue) vs. standard deviation of the OLS estimator (orange).

# 6 Conclusion and Outlook

Our contribution in this paper is two-fold:

Firstly, we have introduced the (1.9) diffusion approximation based on epoched Brownian motion to approximate the behavior of SGD without replacement or reshuffling (SGDo) for the example of linear regression. In an experiment we have justified this approximation by estimating its weak error and comparing it to a deterministic gradient flow approximation.

Secondly, we have used this diffusion approximation as a proxy to predict the convergence properties of SGDo. For linear regression with a constant learning rate we have discovered that the limiting distribution of the diffusion approximation has the same mean as the OLS estimator and its variance converges to the variance of the OLS estimator as the learning rate converges to 0. Further, for the sequence of learning rates $\eta_n = \frac{1}{1+n}$ the limiting distribution has the same mean and the same variance as the OLS estimator.

A direction for future research is to formally state and prove a precise approximation result relating SGDo to (one of) its diffusion approximation(s), either using weak approximation techniques similar to [3], [4], [1] or using strong approximation techniques via a coupling argument, cf. [2].

In general it is then reasonable to allow the diffusion coefficient to be dependent on the state $X$. In this case a discussion of an integral of the type $\int \sigma(X)\, d\hat{W}^T$ is in order. By a careful construction of a filtration one may

interpret this as an integral in the Itô sense. Alternatively, one can treat this integral as a rough integral by lifting $\hat{W}^T$ to a second-order Gaussian rough path.

Finally, we believe that the results can be extended to SGDo processes beyond linear regression using general drift and diffusion coefficients satisfying certain regularity conditions.

# References


[1] S. Ankirchner and S. Perko. Approximating stochastic gradient descent with diffusions: error expansions and impact of learning rate schedules. Oct. 2021. working paper or preprint.

[2] 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.

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

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

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

[6] D. Nagaraj, P. Jain, and P. Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. 97:4703–4711, 09–15 Jun 2019.

[7] O. Shamir. Without-replacement sampling for stochastic gradient methods. 29, 2016.

[8] G. Turinici. The convergence of the stochastic gradient descent (sgd): a self-contained proof. *arXiv preprint arXiv:2103.14350*, 2021.