Journal of Machine Learning Research 24 (2023) 1-48 Submitted 12/21; Published 7/23

A Continuous-time Stochastic Gradient Descent Method for Continuous Data

Kexin Jin KEXINJ@MATH.PRINCETON.EDU
Department of Mathematics
Princeton University
Princeton, NJ 08544-1000, USA

Jonas Latz JONAS.LATZ@MANCHESTER.AC.UK
Department of Mathematics
The University of Manchester
Manchester, M13 9PL, United Kingdom

Chenguang Liu C.LIU-13@TUDELFT.NL
Delft Institute of Applied Mathematics
Technische Universiteit Delft
Delft, 2628 CD, The Netherlands

Carola-Bibiane Schönlieb CBS31@CAM.AC.UK
Department of Applied Mathematics and Theoretical Physics
University of Cambridge
Cambridge, CB3 0WA, United Kingdom

Editor: Andrea Montanari

Abstract

Optimization problems with continuous data appear in, e.g., robust machine learning, functional data analysis, and variational inference. Here, the target function is given as an integral over a family of (continuously) indexed target functions—integrated with respect to a probability measure. Such problems can often be solved by stochastic optimization methods: performing optimization steps with respect to the indexed target function with randomly switched indices. In this work, we study a continuous-time variant of the stochastic gradient descent algorithm for optimization problems with continuous data. This so-called stochastic gradient process consists in a gradient flow minimizing an indexed target function that is coupled with a continuous-time index process determining the index. Index processes are, e.g., reflected diffusions, pure jump processes, or other Lévy processes on compact spaces. Thus, we study multiple sampling patterns for the continuous data space and allow for data simulated or streamed at runtime of the algorithm. We analyze the approximation properties of the stochastic gradient process and study its longtime behavior and ergodicity under constant and decreasing learning rates. We end with illustrating the applicability of the stochastic gradient process in a polynomial regression problem with noisy functional data, as well as in a physics-informed neural network.

Keywords: Stochastic optimization, functional data analysis, robust learning, arbitrary data sources, Markov processes

1. Introduction

The training of a machine learning model is often represented through an optimization problem, where the goal is to calibrate the model's parameters to optimize its goodness-of-fit with respect to training data. The goodness-of-fit is usually quantified through a loss function that sums up the losses from the misrepresentation of every single training data set, see, e.g., Goodfellow et al. (2016) for or, e.g., Hansen (2010); Cam (1990) for similar optimization problems in imaging and statistics. In 'big data' settings, the sum of these loss functions consists of thousands or millions of terms, making classical optimization methods computationally infeasible. Thus, efficiently solving optimization problems of this form has been a focus of machine learning and optimization research in the past decades. Here, methods often build upon the popular stochastic gradient descent method.

Originally, stochastic gradient descent was proposed by Robbins and Monro (1951) to optimize not only sums of loss functions, but also expectations of randomized functions.[1] Of course, a normalized sum is just a special case of an expected value, making stochastic gradient descent available for the kind of training problem described above. Based on stochastic gradient descent ideas, improved algorithms have been proposed for optimizing sums of loss functions, such as Chambolle et al. (2018); Defazio et al. (2014); Duchi et al. (2011). Unfortunately, these methods often specifically target sums of loss functions and can then be infeasible to optimize general expected values of loss functions.

The optimization of expected values of loss functions appears in the presence of countably infinite and continuous data in functional data analysis and non-parametric statistics (e.g., Sinova et al., 2018), physics-informed deep learning (e.g., Raissi et al., 2019), inverse problems (e.g., Bredies and Lorenz, 2018), and continuous data augmentation/adversarial robustness (e.g., Cohen et al., 2019; Shorten and Khoshgoftaar, 2019; Pinto et al., 2017). Some of these problems are usually studied after discretising the data. As hinted above, algorithms for discrete data sometimes deteriorate when approaching the continuum limit, i.e. as the number of data sets goes to infinity. Thus, we prefer studying the continuum case immediately. Finally, we note that 'continuous data' can also refer to optimization under general noise models. Here, expected values are minimized in robust optimization (e.g., Nemirovski et al., 2009), variational Bayesian inference (e.g., Cherief-Abdellatif, 2019), and optimal control (e.g., May et al., 2013). Overall, the optimization of general expected values is a very important task in modern data science, machine learning, and related fields.

In this work, we study stochastic gradient descent for general expected values in a continuous-time framework. We now proceed with the formal introduction of the optimization problem, the stochastic gradient descent algorithm, its continuous-time limits, and current research in this area.

1.1 Problem setting and state of the art

We study optimization problems of the form

$$\min_{\theta \in X} \Phi(\theta) := \int_S f(\theta, y)\pi(\mathrm{d}y), \tag{1}$$

1. Actually, the 'stochastic approximation method' of Robbins and Monro (1951) aims at finding roots of functions that are given as expectations of randomized functions. The method they construct resembles stochastic gradient descent for a least squares loss function.

where $X := \mathbb{R}^K$, S is a Polish space, $f : S \times X \to \mathbb{R}$ is a measurable function that is continuously differentiable in the first variable, and π is a probability measure on S. Moreover, we assume that the integral above always exists. We refer to X as *parameter space*, S as *index set*, Φ as *full target function*, and f as *subsampled target function*. In these optimization problems, it is usually impossible or intractable to evaluate the integral Φ or its gradient $\nabla\Phi$.Hence, traditional optimization algorithms, such as steepest gradient descent or Newton methods are not applicable.

As mentioned above, it is possible to employ stochastic optimization methods, such as the *stochastic gradient descent (SGD)* method, see Kushner and Yin (2003); Robbins and Monro (1951). The stochastic gradient descent method for (1) proceeds through the following discrete-time dynamic that iterates over $n \in \mathbb{N} := \{1, 2, \ldots\}$:

$$\theta_n = \theta_{n-1} - \eta_n \nabla_\theta f(\theta_{n-1}, y_n), \tag{2}$$

where $y_1, y_2, \ldots \sim \pi$ independent and identically distributed (i.i.d.), $(\eta_n)_{n=1}^\infty \in (0, \infty)^{\mathbb{N}}$ is a non-increasing sequence of *learning rates*, and $\theta_0 \in X$ is an appropriate initial value. Hence, SGD is an iterative method that employs only the gradient of the integrand f, but not Φ. SGD converges to the minimizer of Φ, if $\eta_n \to 0$, as $n \to \infty$, sufficiently slowly, and $f(\cdot, y)$ is strongly convex and $\nabla f(\cdot, y)$ is bounded ($y \in S$); see, e.g., Bubeck (2015). SGD is used in practice also for non-convex optimization problems and with constant learning rate. The constant learning rate setting is popular especially due to its regularizing properties; see Ali et al. (2020); Smith et al. (2021).

To understand, improve, and study discrete-time dynamical systems, it is sometimes advantageous to represent them in continuous time, see, e.g. the works by de Wiljes et al. (2018); Kovachki and Stuart (2021); Trillos and Sanz-Alonso (2020). Continuous-time models allow us to concentrate on the underlying dynamics and omit certain numerical considerations. Moreover, they give us natural ways to construct new, efficient algorithms.

The discrete-time dynamic in (2) is sometimes represented through a continuous-time diffusion process, see Ali et al. (2020); Li et al. (2019, 2017); Mandt et al. (2016, 2017); Wojtowytsch (2021):

$$\mathrm{d}\theta_t = -\nabla\Phi(\theta_t)\mathrm{d}t + \sqrt{\eta(t)}\Sigma(\theta_t)^{1/2}\mathrm{d}W_t,$$

where $\Sigma(\theta) = \int (\nabla_\theta f(\theta; y) - \nabla_\theta \Phi(\theta)) \otimes (\nabla_\theta f(\theta; y) - \nabla_\theta \Phi(\theta))\pi(\mathrm{d}y)$, $(W_t)_{t\geq 0}$ is a K-dimensional Brownian motion, and $(\eta(t))_{t\geq 0}$ is an interpolation of the learning rate sequence. While this diffusion approach is suitable to describe the dynamic of the moments of SGD, it does not immediately allow us to construct new stochastic optimization algorithms, as the system depends on the inaccessible $\nabla\Phi$.

A continuous-time representation of stochastic gradient descent that does not depend on $\nabla\Phi$ has recently been proposed by Latz (2021). This work only considers the discrete data case, i.e., S is finite and $\pi := \mathrm{Unif}(S)$. SGD is represented by the *stochastic gradient process* $(\theta_t^\dagger)_{t\geq 0}$. It is defined through the coupled dynamical system

$$\mathrm{d}\theta_t^\dagger = -\nabla_\theta f(\theta_t^\dagger; \mathbf{i}(t))\mathrm{d}t, \tag{3}$$

where $(\mathbf{i}(t))_{t\geq 0}$ is a suitable continuous-time Markov process on S, which we call *index process*. Hence, the process $(\theta_t^\dagger)_{t\geq 0}$ represents gradient flows with respect to the subsampled

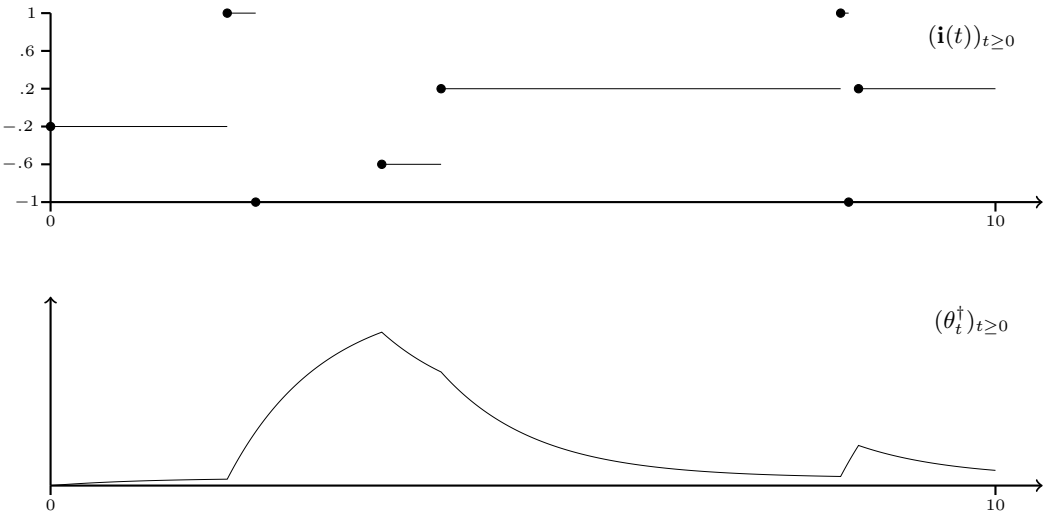

Figure 1: Cartoon of the stochastic gradient process $(\theta_t^\dagger)_{t\geq 0}$ with index process $(\mathbf{i}(t))_{t\geq 0}$ on the discrete index set $S := \{-1, -0.6, \ldots, 1\}$. The index process is a Markov pure jump process on S. The process $(\theta_t^\dagger)_{t\geq 0}$ aims at optimizing the $\mathrm{Unif}(S)$-integral of the subsampled target functional is $f(\theta, y) := \frac{1}{2}(\theta - y^2)^2$ $(\theta \in X := \mathbb{R}, y \in S)$.

target functions that are switched after random waiting times. The random waiting times are controlled by the continuous-time Markov process $(\mathbf{i}(t))_{t\geq 0}$. We show an example of the coupling of exemplary process $(\mathbf{i}(t))_{t\geq 0}$ and $(\theta_t^\dagger)_{t\geq 0}$ in Figure 1. The setting is $S := \{-1, -0.6, \ldots, 1\}$, $\pi := \mathrm{Unif}(S)$, $X := \mathbb{R}$, and $f(\theta, y) := \frac{1}{2}(\theta - y^2)^2$ $(\theta \in X, y \in S)$. There, we see that the sample path of $(\theta_t^\dagger)_{t\geq 0}$ is piecewise smooth, with non-smooth behavior at the jump times of $(\mathbf{i}(t))_{t\geq 0}$.

If the process $(\mathbf{i}(t))_{t\geq 0}$ is homogeneous-in-time, the dynamical system represents a constant learning rate. Inhomogeneous $(\mathbf{i}(t))_{t\geq 0}$ with decreasing mean waiting times, on the other hand, model a decreasing learning rate. Under certain assumptions, the process $(\theta_t^\dagger)_{t\geq 0}$ converges to a unique stationary measure when the learning rate is constant or to the minimizer of Φ when the learning rate decreases.

1.2 This work.

We now briefly introduce the continuous-time stochastic gradient descent methods that we study throughout this work. Then, we summarize our main contributions and give a short paper outline.

In the present work, we aim to generalize the dynamical system (3) to include more general spaces S and probability measures π – studying the more general optimization problems of type (1). We proceed as follows: We define a stationary continuous-time Markov process $(V_t)_{t\geq 0}$ on S that is geometrically ergodic and has π as its stationary measure. This process $(V_t)_{t\geq 0}$ is now our *index process*. Similarly to (3), we then couple $(V_t)_{t\geq 0}$ with the

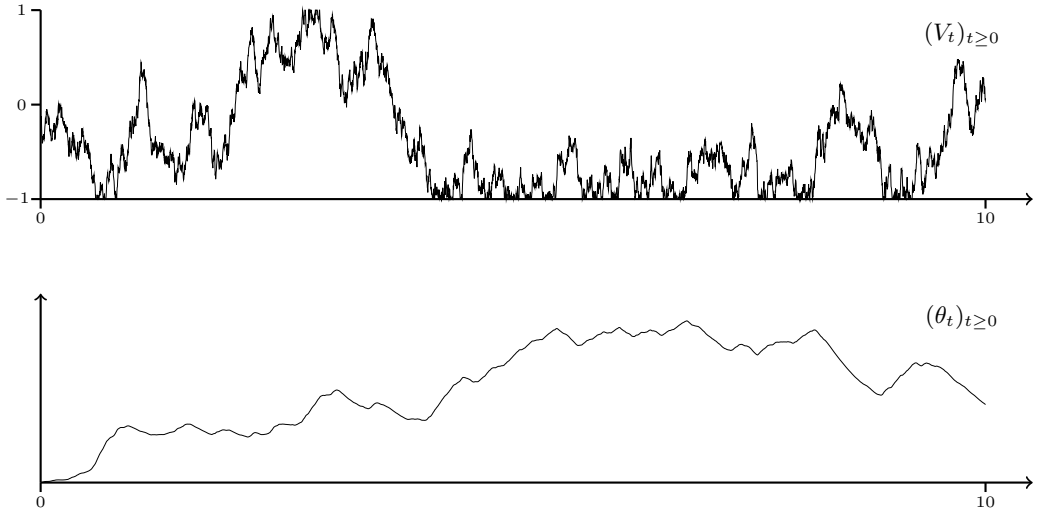

Figure 2: Cartoon of the stochastic gradient process $(\theta_t)_{t\geq 0}$ with index process $(V_t)_{t\geq 0}$, with continuous index set $S := [-1, 1]$. The index process is a reflected Brownian motion on S.

following gradient flow:

$$\mathrm{d}\theta_t = -\nabla_\theta f(\theta_t, V_t)\mathrm{d}t. \tag{4}$$

Note that the index process $(V_t)_{t\geq 0}$ can be considerably more general than the Markov jump processes studied by Latz (2021); we discuss examples below and in Section 2. As the dynamical system (4) contains the discrete version (3) as a special case, we refer to $(\theta_t)_{t\geq 0}$ also as *stochastic gradient process*.

We give an example for $(\theta_t, V_t)_{t\geq 0}$ in Figure 2. There, we consider $S := [-1, 1]$, $\pi := \mathrm{Unif}[-1, 1]$, $X := \mathbb{R}$, and $f(\theta, y) := \frac{1}{2}(\theta - y^2)^2$ $(\theta \in X, y \in S)$. A suitable choice for $(V_t)_{t\geq 0}$ is a reflected Brownian motion on $[-1, 1]$. Although it is coupled with $(V_t)_{t\geq 0}$, the process $(\theta_t)_{t\geq 0}$ appears to be relatively smooth. This may be due to the smoothness of the subsampled target function f. Moreover, we note that the example in Figure 1 is a discretized data version of the example here in Figure 2.

More similarly to the discrete data case (3), one could also choose $(V_t)_{t\geq 0}$ to be a Markov pure jump process on S that has π as a stationary measure. Indeed, the reflected Brownian motion was constructed rather artificially. Sampling from $\mathrm{Unif}[-1, 1]$ is not actually difficult in practice and we just needed a way to find a continuous-time Markov process that is stationary with respect to $\mathrm{Unif}[-1, 1]$. However, there are cases, where one may not be able to sample independently from π. For instance, π could be the measure of interest in a statistical physics simulation or Bayesian inference. In those cases, Markov chain Monte Carlo methods are used to approximate π through a Markov chain stationary with respect to it, see, e.g., Robert and Casella (2004). In other cases, the data might be time series data that is streamed at runtime of the algorithm – a related problem has been studied by Sirignano and Spiliopoulos (2017). Hence, in this work, we also discuss stochastic

optimization in those cases or – more generally – stochastic optimization with respect to data from arbitrary sources.

As the index process $(V_t)_{t\geq 0}$ is stationary, the stochastic gradient process as defined above would, again, represent the situation of a constant learning rate $(\eta_n)_{n=1}^{\infty}$. However, as before we usually cannot hope for convergence to a stationary point if there is not a sense of a decreasing learning rate. Hence, we need to introduce an inhomogeneous variant of $(V_t)_{t\geq 0}$ that represents a decreasing learning rate. We now introduce a way to obtain such a decreasing learning rate in our continuous setting.

We start with the stochastic process $(V_t^{\mathrm{dc}})_{t\geq 0}$ that represents the index process associated to the discrete-time stochastic gradient descent dynamic (2) with constant learning rate parameter $\eta = 1$. This index process is given by

$$V_t^{\mathrm{dc}} = \sum_{n=1}^{\infty} y_n \mathbf{1}[t \in [n-1, n)] = \sum_{n=1}^{\infty} y_n \mathbf{1}[t - n + 1 \in [0, 1)] \qquad (t \geq 0),$$

where $y_1, y_2, \ldots \sim \pi$ i.i.d. and $\mathbf{1}[\cdot]$ is the indicator function: $\mathbf{1}[\text{true}] = 1$ and $\mathbf{1}[\text{false}] = 0$. We now want to turn the process $(V_t^{\mathrm{dc}})_{t\geq 0}$ into the index process $(V_t^{\mathrm{dd}})_{t\geq 0}$ that represents a decreasing learning rate $(\eta_n)_{n=1}^{\infty}$. It is defined through:

$$(V_t^{\mathrm{dd}})_{t\geq 0} = \sum_{n=1}^{\infty} y_n \mathbf{1}\left[t \in [H_{n-1}, H_n)\right] = \sum_{n=1}^{\infty} y_n \mathbf{1}\left[\frac{t - H_{n-1}}{\eta_n} \in [0, 1)\right],$$

where we denote $H_n := \sum_{m=1}^{n} \eta_m$. Hence, we can represent $(V_t^{\mathrm{dd}})_{t\geq 0} := (V_{\beta(t)}^{\mathrm{dc}})_{t\geq 0}$, where $\beta : [0, \infty) \to [0, \infty)$ is given by

$$\beta(t) = \sum_{n=1}^{\infty} \frac{t + n - 1 - H_{n-1}}{\eta_n} \mathbf{1}\left[t \in [H_{n-1}, H_n)\right] \qquad (t \geq 0) \tag{5}$$

is a piecewise linear, non-decreasing function with $\beta(t) \to \infty$, as $t \to \infty$.

Following this idea, we turn our homogeneous index process $(V_t)_{t\geq 0}$ that represents a constant learning rate into an inhomogeneous process with decreasing learning rate using a suitable rescaling function β. In that case, we obtain a stochastic gradient process of type

$$\mathrm{d}\xi_t = -\nabla_{\xi} f(\xi_t, V_{\beta(t)})\mathrm{d}t,$$

which we will use to represent the stochastic gradient descent algorithm with decreasing learning rate. Note that while we require β to satisfy certain conditions that ensure the well-definedness of the dynamical system, it is not strictly necessary for it to be of the form (5). Actually, we later assume that β is smooth.

The main contributions of this work are the following:

- We study stochastic gradient processes for optimization problems of the form (1) with finite, countably infinite, and continuous index sets S.
- We give conditions under which the stochastic gradient process with constant learning rate is well-defined and that it can approximate the full gradient flow $\mathrm{d}\zeta_t = -\nabla\Phi(\zeta_t)\mathrm{d}t$ at any accuracy. In addition, we study the geometric ergodicity of the stochastic gradient process and properties of its stationary measure.

- We study the well-definedness of the stochastic gradient process with decreasing learning rate and give conditions under which the process converges to the minimizer of Φ in the optimization problem (1).

- In numerical experiments, we show the suitability of our stochastic gradient process for (convex) polynomial regression with continuous data and the (non-convex) training of physics-informed neural networks with continuous sampling of function-valued data.

This work is organized as follows. In Section 2, we study the index process $(V_t)_{t\geq 0}$ and give examples for various combinations of index spaces S and probability measures π. Then, in Sections 3 and 4, we analyze the stochastic gradient process with constant and decreasing learning rate, respectively. In Section 5, we review discretization techniques that allow us to turn the continuous dynamical systems into practical optimization algorithms. We employ these techniques in Section 6, where we present numerical experiments regarding polynomial regression and the training of physics-informed neural networks. We end with conclusions and outlook in Section 7.

2. The index process: Feller processes and geometric ergodicity

Before we define the stochastic gradient flow, we introduce and study the class of stochastic processes $(V_t)_{t\geq 0}$ that can be used for the data switching in (4). Moreover, we give an overview of appropriate processes for various measures π. For more background material on (continuous-time) stochastic processes, we refer the reader to the book by Revuz and Yor (2013), the book by Liggett (2010), and other standard literature.

Let $\mathcal{S} = (S, m)$ be a compact Polish space and

$$\Omega = \{\omega : [0, \infty) \to S \mid \omega \text{ is right continuous with left limits.}\}.$$

We consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, (\mathbb{P}_x)_{x\in S})$, where $\mathcal{F}$ is the smallest σ-algebra on Ω such that the mapping $\omega \to \omega(t)$ is measurable for any $t \geq 0$ and the filtration $\mathcal{F}_t$ is right continuous. Let $(V_t)_{t\geq 0}$ be a $(\mathcal{F}_t)_{t\geq 0}$ adapted stochastic process from Ω to S. We assume that $(V_t)_{t\geq 0}$ is Feller with respect to $(\mathcal{F}_t)_{t\geq 0}$. $(\mathbb{P}_x)_{x\in S}$ is a collection of probability measures on Ω such that $\mathbb{P}_x(V_0 = x) = 1$. For any probability measure μ on S, we define

$$\mathbb{P}_\mu(\cdot) := \int_S \mathbb{P}_y(\cdot)\mu(\mathrm{d}y)$$

and denote expectations with respect to $\mathbb{P}_x$ and $\mathbb{P}_\mu$ by $\mathbb{E}_x$ and $\mathbb{E}_\mu$, respectively.

Below we give a set of assumptions on the process $(V_t)_{t\geq 0}$. We need those to ensure that a certain coupling property holds. We comment on these assumptions after stating them.

Assumption 1 *Let $(V_t)_{t\geq 0}$ be a Feller process on $(\Omega, \mathcal{F}, (\mathcal{F}_t))_{t\geq 0}, (\mathbb{P}_x)_{x\in S})$. We assume the following:*

(i) $(V_t)_{t\geq 0}$ admits a unique invariant measure π.

(ii) *For any $x \in S$, there exist a family $(V_t^x)_{t\geq 0}$ and a stationary version $(V_t^\pi)_{t\geq 0}$ defined on the same probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that, $(V_t^x)_{t\geq 0} \overset{d}{=} (V_t)_{t\geq 0}$ in $\mathbb{P}_x$ and $(V_t^\pi)_{t\geq 0} \overset{d}{=} (V_t)_{t\geq 0}$ in $\mathbb{P}_\pi$, i.e. for any $0 \leq t_1 < \cdots < t_n$,*

$$\tilde{\mathbb{P}}(V_{t_1}^x \in A_1, \cdots, V_{t_n}^x \in A_n) = \mathbb{P}_x(V_{t_1} \in A_1, \cdots, V_{t_n} \in A_n),$$

$$\tilde{\mathbb{P}}(V_{t_1}^\pi \in A_1, \cdots, V_{t_n}^\pi \in A_n) = \mathbb{P}_\pi(V_{t_1} \in A_1, \cdots, V_{t_n} \in A_n),$$

where $A_1, \cdots, A_n \in \mathcal{B}(S)$.

(iii) *Let $T^x := \inf\{t \geq 0 \mid V_t^x = V_t^\pi\}$ be a stopping time. There exist constants $C, \delta > 0$ such that for any $t \geq 0$,*

$$\sup_{x\in S} \tilde{\mathbb{P}}(T^x \geq t) \leq C\exp(-\delta t).$$

First, we assume that $(V_t)_{t\geq 0}$ has a stationary measure π. Second, we assume that for the process $(V_t)_{t\geq 0}$ that starts from x with probability 1, we can find a coupled process $(V_t^x)_{t\geq 0}$. Also, given that the process $(V_t)_{t\geq 0}$ starts with its invariant measure π, we can find a stationary version $(V_t^\pi)_{t\geq 0}$ of $(V_t)_{t\geq 0}$. Here, the processes $(V_t^x)_{t\geq 0}$ and $(V_t^\pi)_{t\geq 0}$ are defined on the same probability space. Third, we assume that the processes $(V_t^x)_{t\geq 0}$ and $(V_t^\pi)_{t\geq 0}$ intersect exponentially fast. The exponential rate can be chosen uniformly in x since S is compact. With Assumption 1, we have the following lemma.

Lemma 1 (Geometric Ergodicity) *Under Assumption 1, there exist constants $C, \delta > 0$ such that for any $x \in S$ and $t \geq 0$,*

$$\sup_{A\in\mathcal{B}(S)} |\mathbb{P}_x(V_t \in A) - \pi(A)| \leq C\exp(-\delta t),$$

where $\mathcal{B}(S)$ is the set of all Borel measurable sets of S.

Proof For any given $x \in S$, we construct the following process by coupling $(V_t^x)_{t\geq 0}$ and $(V_t^\pi)_{t\geq 0}$:

$$\tilde{V}_t^x = \begin{cases} V_t^x, & 0 \leq t \leq T^x, \\ V_t^\pi, & t > T^x. \end{cases}$$

By the strong Markov property, $(\tilde{V}_t^x)_{t\geq 0} \overset{d}{=} (V_t^x)_{t\geq 0}$. For any $A \in \mathcal{B}(S)$, notice that

$$\begin{aligned}
&|\mathbb{P}_x(V_t \in A) - \pi(A)| \\
&\quad = |\tilde{\mathbb{P}}(V_t^x \in A) - \tilde{\mathbb{P}}(V_t^\pi \in A)| \\
&\quad = |\tilde{\mathbb{P}}(\tilde{V}_t^x \in A) - \tilde{\mathbb{P}}(V_t^\pi \in A)| \\
&\quad = |\tilde{\mathbb{P}}(\tilde{V}_t^x \in A, \tilde{V}_t^x \neq V_t^\pi) + \tilde{\mathbb{P}}(\tilde{V}_t^x \in A, \tilde{V}_t^x = V_t^\pi) \\
&\qquad\quad - (\tilde{\mathbb{P}}(V_t^\pi \in A, \tilde{V}_t^x \neq V_t^\pi) + \tilde{\mathbb{P}}(V_t^\pi \in A, \tilde{V}_t^x = V_t^\pi))| \\
&\quad = |\tilde{\mathbb{P}}(\tilde{V}_t^x \in A, \tilde{V}_t^x \neq V_t^\pi) - \tilde{\mathbb{P}}(V_t^\pi \in A, \tilde{V}_t^x \neq V_t^\pi)| \\
&\quad \leq 2\tilde{\mathbb{P}}(\tilde{V}_t^x \neq V_t^\pi) \\
&\quad \leq 2\tilde{\mathbb{P}}(T^x \geq t) \leq C\exp(-\delta t).
\end{aligned}$$

From the third assumption in Assumption 1, C and δ are independent of x and this completes the proof. ■

In the lemma above, we have shown geometric ergodicity of $(V_t)_{t\geq 0}$ in the total variation distance. Next, we show that the same rate of convergence of $(V_t)_{t\geq 0}$ holds in the weak topology.

Corollary 2 *Under Assumption 1, there exist constants $C, \delta > 0$ such that for any $h \in \mathcal{C}(S)$, i.e. the set of all continuous function on S, we have*

$$\sup_{x\in S}\left|\mathbb{E}_x[h(V_t)] - \int_S h(y)\pi(\mathrm{d}y)\right| \leq C\left\|h\right\|_\infty \exp(-\delta t)$$

where $\left\|h\right\|_\infty := \sup_{x\in S}|h(x)|$

Proof Rewrite $\mathbb{E}_x[h(V_t)]$ as

$$\mathbb{E}_x[h(V_t)] = \int_S h(y)\mathbb{P}_x(V_t \in \mathrm{d}y).$$

Then, we have

$$\begin{aligned}\left|\mathbb{E}_x[h(V_t)] - \int_S h(y)\pi(\mathrm{d}y)\right| &= \left|\int_S h(y)[\mathbb{P}_x(V_t \in \mathrm{d}y) - \pi(\mathrm{d}y)]\right| \\ &\leq \left\|h\right\|_\infty \left|\mathbb{P}_x(V_t) - \pi\right|(S),\end{aligned}$$

where $\frac{1}{2}\left|\mathbb{P}_x(V_t) - \pi\right|(S)$ is the total variation of the measure $\mathbb{P}_x(V_t) - \pi$. Notice that

$$\left|\mathbb{P}_x(V_t) - \pi\right|(S) = 2\sup_{A\in\mathcal{B}(S)}|\mathbb{P}_x(V_t \in A) - \pi(A)|.$$

By Lemma 1, we have

$$\begin{aligned}\sup_{x\in S}\left|\mathbb{E}_x[h(V_t)] - \int_S h(y)\pi(\mathrm{d}y)\right| &\leq \left\|h\right\|_\infty \sup_{x\in S}\left|\mathbb{P}_x(V_t) - \pi\right|(S) \\ &\leq 2\left\|h\right\|_\infty \sup_{x\in S}\sup_{A\in\mathcal{B}(S)}|\mathbb{P}_x(V_t \in A) - \pi(A)| \\ &\leq C\left\|h\right\|_\infty \exp(-\delta t),\end{aligned}$$

which completes the proof. ■

We now study four examples for processes that satisfy our assumptions: Lévy processes with two-sided reflections on a compact interval, continuous-time Markov processes on finite and countably infinite spaces, and processes on rectangular sets with independent coordinates.

2.1 Example 1: Lévy processes with two-sided reflection

For any $b > 0$, we say a triplet $((X_t)_{t\geq 0}, (L_t)_{t\geq 0}, (U_t)_{t\geq 0})$ is a solution to the Skorokhod problem of the Lévy process $(X_t)_{t\geq 0}$ on the space $S := [0, b]$ if for all $t \geq 0$,

$$V_t = X_t + L_t - U_t, \tag{6}$$

where $(L_t)_{t\geq 0},\ (U_t)_{t\geq 0}$ are non-decreasing right continuous processes such that

$$\int_0^\infty V_t \mathrm{d}L_t = \int_0^\infty (b - V_t)\mathrm{d}U_t = 0.$$

In other words, $(L_t)_{t\geq 0}$ and $(U_t)_{t\geq 0}$ can only increase when $(V_t)_{t\geq 0}$ is at the lower boundary 0 or the upper boundary b. From Andersen et al. (2015, Proposition 5.1), we immediately see that the process $(V_t)_{t\geq 0}$ in (6) satisfies Assumption 1. The geometric ergodicity follows from Andersen et al. (2015, Remark 5.3). As an example, the standard Brownian Motion (BM) reflected at 0 and 1 can be written as

$$V_t = B_t + \tilde{L}_t^0 - \tilde{L}_t^1,$$

where $(B_t)_{t\geq 0}$ is a standard BM and $(\tilde{L}_t^a)_{t\geq 0}$ is the symmetric local time of $(V_t)_{t\geq 0}$ at $a \in \{0, 1\}$. Intuitively, a local time describes the time spent at a given point of a continuous stochastic process. The formal definition of symmetric local time of continuous semimartingales can be found, for example, in Revuz and Yor (2013, Chapter VI). For the optimization problem (1) with $S = [0, 1]$ and π being the uniform measure on S, the corresponding stochastic process in (4) can be chosen to be this Brownian Motion with two-sided refection since its invariant measure is the uniform measure on $[0, 1]$. To see this, we employ Andersen et al. (2015, Theorem 5.4) and have

$$\pi([x, 1]) = \mathbb{P}(B_{\tau_x \wedge \tau_{x-1}} = x) = \mathbb{P}(\tau_x < \tau_{x-1}) = 1 - x \qquad (x \in [0, 1]),$$

where $\tau_a = \inf\{t \geq 0 | B_t = a\}$.

2.2 Example 2: Continuous-time Markov processes with finite states

We consider a continuous-time Markov process V_t on state space $I = \{1, 2, \ldots, N\}$ with transition rate matrix

$$\mathbf{A}_N = \mathbf{\Lambda}_N - N\lambda \mathbf{I}_N,$$

where $\lambda > 0$, $\mathbf{\Lambda}_N$ is an $N \times N$ matrix the entries of which are all equal to λ, and $\mathbf{I}_N$ is the identity matrix. From Latz (2021), we know that the transition probability is given by

$$\mathbb{P}(V_{t+s} = i | V_s = j) = \frac{1 - \exp(-\lambda N t)}{N} + \exp(-\lambda N t)\mathbf{1}[i = j].$$

The invariant measure π of V_t is the uniform measure on I, i.e. $\pi(i) = 1/N$ for $i \in \{1, ..., N\}$. To see that $(V_t)_{t\geq 0}$ satisfies the rest of Assumption 1, consider a stationary version $(\hat{V}_t)_{t\geq 0}$ that is independent of $(V_t)_{t\geq 0}$. Let $V_0 = 1$. We define $T = \inf\{t \geq 0, V_t = \hat{V}_t\}$. Moreover,

for $i,\ j \in \mathbb{N}$, we denote the i-th and j-th jump time of $(V_t)_{t\geq 0}$ and $(\hat{V}_t)_{t\geq 0}$ by T_i and $\hat{T}_j$, respectively. Then we have

$$\mathbb{P}(T=0) = \mathbb{P}(V_0 = 1, \hat{V}_0 = 1) = \mathbb{P}(V_0 = 1)\mathbb{P}(\hat{V}_0 = 1) = \frac{1}{N}.$$

For any $i,\ j \in \mathbb{N}$, since T_i and $\hat{T}_j$ are independent, $\mathbb{P}(T_i = \hat{T}_j) = 0$. Let $(Y_t)_{t\geq 0} = ((V_t, \hat{V}_t))_{t\geq 0}$ be a Markov process on $I \times I$ with transition probability:

$$\begin{aligned}\mathbb{P}(Y_{t+s} = (i,j)|Y_s = (i_0, j_0)) = \frac{1-\exp(-2\lambda N t)}{2N}\Big(\mathbf{1}[i = i_0] + \mathbf{1}[j = j_0]\Big)\\ + \exp(-2\lambda N t)\mathbf{1}[(i,j) = (i_0, j_0)].\end{aligned}$$

Thus, T is the first time when $(Y_t)_{t\geq 0}$ hits $\{(i,i)|i = 1, ..., N\}$. Let the n-th jump time of $(Y_t)_{t\geq 0}$ be τ_n, for $t > 0$, we have

$$\begin{aligned}\mathbb{P}(T \geq t) =& \sum_{n\geq 1} \mathbb{P}(T = \tau_n, \tau_n \geq t)\\ =& \sum_{n\geq 1} \exp(-2(N-1)n\lambda t)\frac{N-1}{N}\frac{1}{N-1}\Big(\frac{N-2}{N-1}\Big)^{n-1}\\ \leq& C\exp(-2(N-1)\lambda t),\end{aligned}$$

where the second equality follows from

$$\mathbb{P}(T = \tau_n) = \frac{N-1}{N}\frac{1}{N-1}\Big(\frac{N-2}{N-1}\Big)^{n-1}$$

since there are $2N-4$ states available for the next jump. Thus, $(V_t)_{t\geq 0}$ satisfies Assumption 1.

2.3 Example 3: Continuous-time Markov processes with countable states

We consider a continuous-time Markov process $(V_t)_{t\geq 0}$ on state space $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ with exponential jump times. At time t, if $V_t \in \mathbb{N}$, it jumps to 0 with probability 1 at the next jump time. Otherwise, if $V_t = 0$, it jumps to i with probability $1/2^i$. It is easy to verify that the invariant measure π of $(V_t)_{t\geq 0}$ is $\pi(\{i\}) = 1/2^{i+1}$. One may consider $\mathbb{N}$ as one state and view $(V_t)_{t\geq 0}$ as a Markov process with two states.

To verify that $(V_t)_{t\geq 0}$ satisfies the rest of Assumption 1, similarly to the previous example, we consider a stationary version $(\hat{V}_t)_{t\geq 0}$ that is independent of $(V_t)_{t\geq 0}$. Let $V_0 = 0$ and $T = \inf\{t \geq 0, V_t = \hat{V}_t = 0\}$. For $i,\ j \in \mathbb{N}_0$, we denote the i-th and j-th jump time of $(V_t)_{t\geq 0}$ and $(\hat{V}_t)_{t\geq 0}$ by T_i and $\hat{T}_j$, respectively. Then we have

$$\mathbb{P}(T=0) = \mathbb{P}(V_0 = 0, \hat{V}_0 = 0) = \mathbb{P}(V_0 = 0)\mathbb{P}(\hat{V}_0 = 0) = \frac{1}{2}.$$

For any $i,\ j \in \mathbb{N}_0$, since T_i and $\hat{T}_j$ are independent, $\mathbb{P}(T_i = \hat{T}_j) = 0$. Let $(Y_t)_{t\geq 0} = ((V_t, \hat{V}_t))_{t\geq 0}$ be a Markov process on $\mathbb{N}_0 \times \mathbb{N}_0$. Notice that T is the first time when $(Y_t)_{t\geq 0}$

hits $(0,0)$. Let the n-th jump time of $(Y_t)_{t\geq 0}$ be τ_n, for $t>0$, we have

$$\begin{aligned}\mathbb{P}(T\geq t) &= \sum_{n\geq 1}\mathbb{P}(T=\tau_n,\tau_n\geq t)\\ &= \sum_{k\geq 0}\mathbb{P}(T=\tau_{2k+1},\tau_{2k+1}\geq t)\\ &= \sum_{k\geq 0}\exp(-2(2k+1)t)\frac{1}{2^{k+2}}\leq \exp(-t),\end{aligned}$$

where the second and the third equality follows from $\mathbb{P}(T=\tau_{2k+1})=2^{-k-1}$ and $\mathbb{P}(T=\tau_{2k})=0$ for $k\geq 1$. Since $\inf\{t\geq 0, V_t=\hat{V}_t\}$ is upper bounded by T, $(V_t)_{t\geq 0}$ satisfies Assumption 1.

2.4 Example 4: Multidimensional processes

For multidimensional processes, Assumption 1 is satisfied if each component satisfies Assumption 1 and all components are mutually independent. We illustrate this by discussing the 2-dimensional case – higher-dimensional processes can be constructed inductively. Multidimensional processes arise, e.g., when the underlying space S is multidimensional. They also arise when S is one-dimensional, but we run multiple processes in parallel to obtain a mini-batch SGD instead of single-draw SGD.

Let (S^1,m^1) and (S^2,m^2) be two compact Polish spaces. We consider the probability triples $(\Omega^1,(\mathcal{F}^1_t))_{t\geq 0},(\mathbb{P}^1_a)_{a\in S^1})$ and $(\Omega^2,(\mathcal{F}^2_t))_{t\geq 0},(\mathbb{P}^2_b)_{b\in S^2})$ with $\mathbb{P}^1_a((V^1_0=a)=\mathbb{P}^2_b(V^2_0=b)=1$. Let $(V^1_t)_{t\geq 0}$ and $(V^2_t)_{t\geq 0}$ be $(\mathcal{F}^1_t))_{t\geq 0}$ and $(\mathcal{F}^2_t))_{t\geq 0}$ adapted and from Ω^1 to S^1 and Ω^2 to S^2 respectively. In the following proposition, we construct a 2-dimensional process $(V^1_t,V^2_t)_{t\geq 0}$ from $\Omega^1\times\Omega^2$ to $(S^1\times S^2,m^1+m^2)$ with a family of probability measures $(\mathbb{P}_{(a,b)})_{(a,b)\in S^1\times S^2}$ such that $\mathbb{P}_{(a,b)}(A\times B)=\mathbb{P}^1_a(A)\mathbb{P}^2_b(B)$ for $A\in\mathcal{F}^1$ and $B\in\mathcal{F}^2$.

We now show that the joint process $(V^1_t,V^2_t)_{t\geq 0}$ is Feller and satisfies Assumption 1, if the marginals do.

Proposition 1 *Let $(V^1_t)_{t\geq 0}$ and $(V^2_t)_{t\geq 0}$ be càdlàg and Feller with respect to $(\mathcal{F}^1_t)_{t\geq 0}$ and $(\mathcal{F}^2_t)_{t\geq 0}$, respectively, and satisfy Assumption 1 with probability $(\mathbb{P}^1_a)_{a\in S^2}$ and $(\mathbb{P}^2_b)_{b\in S^2}$, respectively. Then $(V^1_t,V^2_t)_{t\geq 0}$ is also càdlàg and Feller with respect to $\sigma(\mathcal{F}^1_t\times\mathcal{F}^2_t)_{t\geq 0}$ and satisfies Assumption 1 with $(\mathbb{P}_{(a,b)})_{(a,b)\in S^1\times S^2}$.*

Proof It is obvious that the process $(V^1_t,V^2_t)_{t\geq 0}$ is càdlàg and Markovian. To verify the Feller property, we show that for any continuous function F on $S^1\times S^2$, $\mathbb{E}_{(x,y)}[F(V^1_t,V^2_t)]$ is continuous in (x,y). We shall prove this by showing this property for separable F and approximate general continuous functions using this special case. Let f and g be continuous functions on S^1 and S^2 respectively, then we have

$$\mathbb{E}_{(x,y)}[f(V^1_t)g(V^2_t)]=\mathbb{E}^1_x[f(V^1_t)]\mathbb{E}^2_y[g(V^2_t)], \tag{7}$$

which implies $\mathbb{E}_{(x,y)}[f(V^1_t)g(V^2_t)]$ is continuous in (x,y) since $(V^1_t)_{t\geq 0}$ and $(V^2_t)_{t\geq 0}$ are Feller. By the Stone–Weierstrass theorem, for any $k\geq 1$, any continuous function F on $S^1\times S^2$

can be approximated as the following,

$$\sup_{(x,y)\in S^1\times S^2}\left|F(x,y)-\sum_{i=1}^{n_k} f_i^k(x)g_i^k(y)\right|\leq \frac{1}{k}$$

where f_i^k and g_i^k are continuous. From (7), this implies $\mathbb{E}_{(x,y)}[F(V_t^1,V_t^2)]$ is continuous on $S^1\times S^2$.

Next, we prove that $(V_t^1,V_t^2)_{t\geq 0}$ satisfies Assumption 1. Let π^1 and π^2 be the invariant measures of $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$, respectively. Then $\pi^1\times\pi^2$ is the invariant measure of $(V_t^1,V_t^2)_{t\geq 0}$ since $(V_t^1)_{t\geq 0}$ and $(V_t^2)_{t\geq 0}$ are independent. From Assumption 1, we know there exist $(\tilde{\Omega}^1,\tilde{\mathcal{F}}^1,\tilde{\mathbb{P}}^1)$, $(\tilde{\Omega}^2,\tilde{\mathcal{F}}^2,\tilde{\mathbb{P}}^2)$, such that for any $a\in S^1$ and $b\in S^2$, $(V_t^{1,a})_{t\geq 0}\overset{d}{=}(V_t^1)_{t\geq 0}$ in $\mathbb{P}_a^1$ and $(V_t^{2,b})_{t\geq 0}\overset{d}{=}(V_t^2)_{t\geq 0}$ in $\mathbb{P}_b^2$. We define $\tilde{\mathbb{P}}$ on $\tilde{\Omega}^1\times\tilde{\Omega}^2$ such that

$$\tilde{\mathbb{P}}(A\times B)=\tilde{\mathbb{P}}^1(A)\tilde{\mathbb{P}}^2(B),\quad (A\in\tilde{\mathcal{F}}^1,\ B\in\tilde{\mathcal{F}}^2).$$

Then we have that $((V_t^{1,a})_{t\geq 0})_{a\in S^1}$ and $(V_t^{1,\pi^1})_{t\geq 0}$ are independent of $((V_t^{2,b})_{t\geq 0})_{b\in S^2}$ and $(V_t^{2,\pi^2})_{t\geq 0}$ under $\tilde{\mathbb{P}}$. Similar to the proof of Lemma 1, we construct the following processes by the coupling method:

$$\tilde{V}_t^{1,a}=\begin{cases}V_t^{1,a}, & 0\leq t\leq T^{1,a},\\ V_t^{1,\pi^1}, & t>T^{1,a},\end{cases}$$

and

$$\tilde{V}_t^{2,b}=\begin{cases}V_t^{2,b}, & 0\leq t\leq T^{2,b},\\ V_t^{2,\pi^2}, & t>T^{2,b}.\end{cases}$$

Then the distribution of $(\tilde{V}_t^{1,a},\tilde{V}_t^{2,b})_{t\geq 0}$ under $\tilde{\mathbb{P}}$ is the same as the distribution of $(V_t^1,V_t^2)_{t\geq 0}$ under $\mathbb{P}_{(a,b)}$; the distribution of $(V_t^{1,\pi^1},V_t^{2,\pi^2})_{t\geq 0}$ under $\tilde{\mathbb{P}}$ is the same as the distribution of $(V_t^1,V_t^2)_{t\geq 0}$ under $\mathbb{P}_{\pi^1\times\pi^2}$. Moreover, $(\tilde{V}_t^{1,a},\tilde{V}_t^{2,b})_{t\geq 0}$ intersects the invariant state $(V_t^{1,\pi^1},V_t^{2,\pi^2})_{t\geq 0}$ at time $T^{1,a}\vee T^{1,b}$. For any $(a,b)\in S^1\times S^2$,

$$\tilde{\mathbb{P}}(T^{1,a}\vee T^{1,b}\geq t)\leq\tilde{\mathbb{P}}(T^{1,a}\geq t)+\tilde{\mathbb{P}}(T^{1,b}\geq t)\leq C\exp(-\delta t).$$

■

We have now discussed various index processes and their properties.We are ready to move on to study the stochastic gradient process.

3. Stochastic gradient processes with constant learning rate

We now define and study the stochastic gradient process with constant learning rate. Here, the switching between data sets is performed in a homogeneous-in-time way. Hence, it models the discrete-time stochastic gradient descent algorithm when employed with a constant learning rate. Although, one can usually not hope to converge to the minimizer of the target functional in this case, this setting is popular in practice.

To obtain the stochastic gradient process with constant learning rate, we will couple the gradient flow (4) with the an appropriate process $(V_{t/\varepsilon})_{t\geq 0}$. Here, $(V_t)_{t\geq 0}$ is a Feller process introduced in Section 2 and $\varepsilon > 0$ is a scaling parameter that allows us to uniformly control a switching rate parameter. To define the stochastic process associated with this stochastic gradient descent problem, we first introduce the following assumptions that guarantee the existence and uniqueness of the solution of the associated dynamical system. After its formal definition and the proof of well-definedness, we move on to the analysis of the process. Indeed, we show that the process approximates the full gradient flow (9), as $\varepsilon \downarrow 0$. Moreover, we show that the process has a unique stationary measure to which it converges in the longtime limit at geometric speed.

We commence with regularity properties of the subsampled target function f that are necessary to show the well-definedness of the stochastic gradient process.

Assumption 2 *Let $f(\theta, y) \in \mathcal{C}^2(\mathbb{R}^K \times S, \mathbb{R})$.*

1. *$\nabla_\theta f$, $H_\theta f$ are continuous,*
2. *$\nabla_\theta f(\theta, y)$ is Lipschitz in x and the Lipschitz constant is uniform for $y \in S$, and*
3. *$f(\theta, \cdot)$ and $\nabla_\theta f$ are integrable w.r.t to the probability measure $\pi(\cdot)$, for $\theta \in \mathbb{R}^K$.*

Now, we move on to the formal definition of the stochastic gradient process.

Definition 1 *For $\varepsilon > 0$, the* stochastic gradient process with constant learning rate (SGPC) *is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}\theta_t^\varepsilon = -\nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\mathrm{d}t, \\ \theta_0^\varepsilon = \theta_0, \end{cases} \tag{8}$$

where f satisfies Assumption 2 and $(V_t)_{t\geq 0}$ is a Feller process that satisfies Assumption 1.

Given these two assumptions, we can indeed show that SGPC is a well-defined Markov process.

Proposition 2 *Let Assumptions 1 and 2 hold. Then, equation (8) has a unique strong solution, i.e. the solution $(\theta_t^\varepsilon)_{t\geq 0}$ is measurable with respect to $\mathcal{F}_t^\varepsilon := \mathcal{F}_{t/\varepsilon}$ for any $t \geq 0$. For $y \in S$, $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$ is a Markov process under $\mathbb{P}_y$ with respect to $(\mathcal{F}_t^\varepsilon)_{t\geq 0}$.*

Proof The existence and the uniqueness of the strong solution to the equation (8) can be found in Kushner (1990, Chapter 2, Theorem 4.1). To prove the Markov property, we define the operator $(Q_t^\varepsilon)_{t\geq 0}$ such that

$$Q_t^\varepsilon h(x, y) := \mathbb{E}_y[h(\theta_t^\varepsilon, V_{t/\varepsilon})|\theta_0^\varepsilon = x],$$

for any function h bounded and measurable on $\mathbb{R}^K \times S$. For any $s, t \geq 0$, we want to show

$$\mathbb{E}[h(\theta_{t+s}^\varepsilon, V_{(t+s)/\varepsilon})|\mathcal{F}_s^\varepsilon] = Q_t^\varepsilon h(\theta_s^\varepsilon, V_{s/\varepsilon}).$$

We set $\hat{\theta}_t^\varepsilon := \theta_{t+s}^\varepsilon$, $\hat{\mathcal{F}}_t := \mathcal{F}_{t+s}^\varepsilon$, $\hat{V}_{t/\varepsilon} := V_{(t+s)/\varepsilon}$. Since

$$\theta_{t+s}^\varepsilon = \theta_s^\varepsilon - \int_s^{t+s} \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\mathrm{d}m,$$

we have

$$\hat{\theta}_t^\varepsilon = \hat{\theta}_0^\varepsilon - \int_0^t \nabla_\theta f(\hat{\theta}_m^\varepsilon, \hat{V}_{m/\varepsilon})\mathrm{d}m.$$

Hence $\hat{\theta}_t^\varepsilon$ is the solution of equation (8) with $\hat{\theta}_0^\varepsilon = \theta_s^\varepsilon$ and $\hat{V}_0 = V_{s/\varepsilon}$. Moreover,

$$\begin{aligned}\mathbb{E}[h(\theta_{t+s}^\varepsilon, V_{(t+s)/\varepsilon})|\mathcal{F}_s^\varepsilon] &= \mathbb{E}[h(\hat{\theta}_t^\varepsilon, \hat{V}_{t/\varepsilon})|\hat{\theta}_0^\varepsilon = \theta_s^\varepsilon, \hat{V}_0 = V_{s/\varepsilon}] \\ &= \mathbb{E}_{\hat{V}_0}[h(\hat{\theta}_t^\varepsilon, \hat{V}_{t/\varepsilon})|\hat{\theta}_0^\varepsilon = \theta_s^\varepsilon] \\ &= Q_t^\varepsilon h(\theta_s^\varepsilon, V_{s/\varepsilon}),\end{aligned}$$

where the second equality and third equality follow from the homogeneous Markov property of $(V_t^\varepsilon)_{t\geq 0}$. ■

3.1 Approximation of the full gradient flow

We now let $\varepsilon \to 0$ and study the limiting behavior of SGPC. Indeed, we aim to show that here the SGPC converges to the *full gradient flow*

$$\mathrm{d}\zeta_t = -\Big[\int_S \nabla_\zeta f(\zeta_t, v)\pi(\mathrm{d}v)\Big]\mathrm{d}t. \tag{9}$$

We study this topic for two reasons: First, we aim to understand the interdependence of $(V_t)_{t\geq 0}$ and $(\theta_t^\varepsilon)_{t\geq 0}$. Second, we understand SGPC as an approximation to the full gradient flow (9), as motivated in the introduction. Hence, we should show that SGPC can approximate the full gradient flow at any accuracy.

We now denote $g(\cdot) := \int_S \nabla_\zeta f(\cdot, v)\pi(\mathrm{d}v) \in \mathcal{C}^1(\mathbb{R}^K, \mathbb{R}^K)$. Then, we can define $(\zeta_t)_{t\geq 0}$ through the dynamical system $\mathrm{d}\zeta_t = -g(\zeta_t)\mathrm{d}t$. Moreover, let $\mathcal{C}([0,\infty) : \mathbb{R}^K)$ be the space of continuous functions from $[0,\infty)$ to $\mathbb{R}^K$ equipped with the distance

$$\rho\Big((\varphi_t)_{t\geq 0}, (\varphi_t')_{t\geq 0}\Big) := \int_0^\infty \exp(-t)(1 \wedge \sup_{0\leq s\leq t} \big\|\varphi_s - \varphi_s'\big\|)\mathrm{d}t,$$

where $(\varphi_t)_{t\geq 0}, (\varphi_t')_{t\geq 0} \in \mathcal{C}([0,\infty) : \mathbb{R}^K)$. We study the weak limit of the system (8) as $\varepsilon \to 0$. Similar problems have been discussed in, for example, Kushner (1990) and Kushner (1984).

Theorem 3 *Let $\theta_0^\varepsilon = \theta_0$ and $\zeta_0 = \theta_0$. Moreover, let $(\theta_t^\varepsilon)_{t\geq 0}$ and $(\zeta_t)_{t\geq 0}$ solve (8) and (9), respectively. Then $(\theta_t^\varepsilon)_{t\geq 0}$ under $\mathbb{P}_\pi$ converges weakly to $(\zeta_t)_{t\geq 0}$ in $\mathcal{C}([0,\infty) : \mathbb{R}^K)$ as $\varepsilon \to 0$, i.e. for any bounded continuous function $F : \mathcal{C}([0,\infty) : \mathbb{R}^K) \to \mathbb{R}$, we have*

$$\mathbb{E}_\pi[F\big((\theta_t^\varepsilon)_{t\geq 0}\big)] \to \mathbb{E}_\pi[F\big((\zeta_t)_{t\geq 0}\big)] = F\big((\zeta_t)_{t\geq 0}\big).$$

Proof We first verify that $(\theta_t^\varepsilon)_{t\geq 0}$ is tight by checking:

1. $\sup_{0<\varepsilon<1} \|\theta_0^\varepsilon\| < +\infty$;
2. *For any fixed $T > 0$,* $\lim_{\delta\to 0} \sup_{0<\varepsilon<1} \sup_{s,t\in[0,T],|s-t|\leq\delta} \|\theta_t^\varepsilon - \theta_s^\varepsilon\| \to 0.$

The first condition follows from $\theta_0^\varepsilon = \theta_0$. For the second condition, by Assumption 2, let $C_0 = \sup_{y\in S} \|\nabla_\theta f(0,y)\|$ and L_f be the Lipschitz constant of $\nabla_\theta f(\cdot, y)$, we have

$$\begin{aligned}\frac{\mathrm{d}\left\|\theta_t^\varepsilon\right\|^2}{\mathrm{d}t} &= -2\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\right\rangle \\ &= -2\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon}) - \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle - 2\left\langle \theta_t^\varepsilon, \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle \\ &\le 2L_f\left\|\theta_t^\varepsilon\right\|^2 + 2C_0\left\|\theta_t^\varepsilon\right\| \\ &\le 2L_f\left\|\theta_t^\varepsilon\right\|^2 + \left\|\theta_t^\varepsilon\right\|^2 + C_0^2 \\ &= (2L_f+1)\left\|\theta_t^\varepsilon\right\|^2 + C_0^2.\end{aligned}$$

By Grönwall's inequality,

$$\left\|\theta_t^\varepsilon\right\|^2 \le (\|\theta_0\|^2 + C_0^2)e^{(2L_f+1)t}. \tag{10}$$

Therefore, θ_t^ε is bounded on any finite time interval. For any fixed $T > 0$, let

$$C_{T,f,\theta_0} = \sup_{\|x\|\le(\|\theta_0\|^2+C_0^2)e^{(2L_f+1)T},\ y\in S} \|\nabla_\theta f(x,y)\|.$$

Then for any $s, t \in [0,T]$,

$$\|\theta_t^\varepsilon - \theta_s^\varepsilon\| \le \int_s^t \left\|\nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\right\| \mathrm{d}m \le C_{T,f,\theta_0}|t-s|.$$

Hence, $(\theta_t^\varepsilon)_{t\ge0}$ is tight in $\mathcal{C}([0,\infty):\mathbb{R}^K)$. By Prokhorov's theorem, let $(\theta_t)_{t\ge0}$ be a weak limit of $(\theta_t^\varepsilon)_{t\ge0}$. We shall verify that $(\theta_t)_{t\ge0}$ satisfies equation (9), which is equivalent to show that for any bounded differentiable function φ, h

$$\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t) - \varphi(\theta_s) + \int_s^t \langle\nabla_\theta\varphi(\theta_m), g(\theta_m)\rangle\,\mathrm{d}m\Big)h\Big((\theta_{t_i})_{i=1,\dots,n}\Big)\Big] = 0,$$

$\forall\ 0 \le t_1 < \cdots < t_n \le s$. The case $t = 0$ is obvious. Since $(\theta_t^\varepsilon)_{t\ge0}$ is a strong solution to equation (8), for any $0 \le s < t$,

$$\varphi(\theta_t^\varepsilon) = \varphi(\theta_s^\varepsilon) - \int_s^t \left\langle\nabla_\theta\varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\right\rangle \mathrm{d}m. \tag{11}$$

Hence, we have

$$\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t^\varepsilon) - \varphi(\theta_s^\varepsilon) + \int_s^t \left\langle\nabla_\theta\varphi(\theta_s^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\right\rangle \mathrm{d}m\Big)h\Big((\theta_{t_i}^\varepsilon)_{i=1,\dots,n}\Big)\Big] = 0,$$

Moreover, when $\varepsilon \to 0$,

$$\mathbb{E}_\pi\Big[\Big(\varphi(\theta_t^\varepsilon) - \varphi(\theta_s^\varepsilon)\Big)h\Big((\theta_{t_i}^\varepsilon)_{i=1,\dots,n}\Big)\Big] \to \mathbb{E}_\pi\Big[\Big(\varphi(\theta_t) - \varphi(\theta_s)\Big)h\Big((\theta_{t_i})_{i=1,\dots,n}\Big)\Big].$$

Hence, all we need to show is the following

$$\mathbb{E}_\pi\Big[\Big(\int_s^t \left\langle\nabla_\theta\varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\right\rangle \mathrm{d}m - \int_s^t \langle\nabla_\theta\varphi(\theta_m^\varepsilon), g(\theta_m^\varepsilon)\rangle\,\mathrm{d}m\Big)h\Big((\theta_{t_i}^\varepsilon)_{i=1,\dots,n}\Big)\Big] \to 0, \tag{12}$$

which is equivalent to prove that

$$\mathbb{E}_\pi\Big[\int_s^t \big\langle \nabla_\theta\varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\rangle \,\mathrm{d}m - \int_s^t \langle \nabla_\theta\varphi(\theta_m^\varepsilon), g(\theta_m^\varepsilon)\rangle \,\mathrm{d}m \Big| \mathcal{F}_s^\varepsilon\Big] \to 0. \tag{13}$$

Let $\tilde{\varepsilon} := 1/[1/\sqrt{\varepsilon}]$, where $[x]$ is the greatest integer less than or equal to x. Then we have the following decomposition

$$\begin{aligned}
&\mathbb{E}_\pi\Big[\int_s^t \big\langle \nabla_\theta\varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\rangle \,\mathrm{d}m - \int_s^t \langle \nabla_\theta\varphi(\theta_m^\varepsilon), g(\theta_m^\varepsilon)\rangle \,\mathrm{d}m \Big| \mathcal{F}_s^\varepsilon\Big] \\
=&\tilde{\varepsilon}\sum_{i=0}^{1/\tilde{\varepsilon}} \tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_{s+i(t-s)\tilde{\varepsilon}}^{s+(i+1)(t-s)\tilde{\varepsilon}} \big\langle \nabla_\theta\varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon}) - g(\theta_m^\varepsilon)\big\rangle \,\mathrm{d}m \Big| \mathcal{F}_s^\varepsilon\Big] \\
=&\tilde{\varepsilon}\sum_{i=0}^{1/\tilde{\varepsilon}} \mathbb{E}_\pi\Big[\tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_{s+i(t-s)\tilde{\varepsilon}}^{s+(i+1)(t-s)\tilde{\varepsilon}} \big\langle \nabla_\theta\varphi(\theta_m^\varepsilon), \nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon}) - g(\theta_m^\varepsilon)\big\rangle \,\mathrm{d}m \Big| \mathcal{F}_{s+i(t-s)\tilde{\varepsilon}}^\varepsilon\Big]\Big| \mathcal{F}_s^\varepsilon\Big].
\end{aligned}$$

We claim that as $\varepsilon \to 0$,

$$\sup_{0\le r<t} \tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_m^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big] \to 0, \tag{14}$$

where $G(x,y) := \langle \nabla_\theta\varphi(x), \nabla_\theta f(x,y) - g(x)\rangle$. Notice that for any fixed $t > 0$, $(\theta_s^\varepsilon)_{0\le s\le t}$ is uniformly equicontinuous. Hence, we have

$$\begin{aligned}
\sup_{0\le r\le t}\sup_{r\le m\le r+(t-s)\tilde{\varepsilon}} \|\theta_m^\varepsilon - \theta_r^\varepsilon\| &= \sup_{0\le r\le t}\sup_{r\le m\le r+(t-s)\tilde{\varepsilon}} \int_r^{r+(t-s)\tilde{\varepsilon}} \big\|\nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\| \,\mathrm{d}m \\
&\le \tilde{\varepsilon}\sup_{0\le m\le t} \big\|\nabla_\theta f(\theta_m^\varepsilon, V_{m/\varepsilon})\big\|.
\end{aligned}$$

Therefore, as $\varepsilon \to 0$,

$$\sup_{0\le r<t}\Big|\tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_m^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big] - \tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_r^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big]\Big| \to 0.$$

Hence, (14) is equivalent to

$$\sup_{0\le r<t}\Big|\tilde{\varepsilon}^{-1}\mathbb{E}_\pi\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_r^\varepsilon, V_{m/\varepsilon})\mathrm{d}m \Big| \mathcal{F}_r^\varepsilon\Big]\Big| \to 0. \tag{15}$$

By Corollary 2,

$$
\begin{aligned}
&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\mathbb{E}_{\pi}\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(\theta_r^{\varepsilon},V_{m/\varepsilon})\mathrm{d}m\Big|\mathcal{F}_r^{\varepsilon}\Big]\right|\\
=&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} G(x,V_{(m-r)/\varepsilon})\mathrm{d}m\Big]\right|\\
=&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\int_r^{r+(t-s)\tilde{\varepsilon}} \big\langle\nabla_\theta\varphi(x),\nabla_\theta f(x,V_{(m-r)/\varepsilon})-g(x)\big\rangle\,\mathrm{d}m\Big]\right|\\
=&\sup_{0\le r<t}\left|\tilde{\varepsilon}^{-1}\int_r^{r+(t-s)\tilde{\varepsilon}} \mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\big\langle\nabla_\theta\varphi(x),\nabla_\theta f(x,V_{(m-r)/\varepsilon})-g(x)\big\rangle\Big]\mathrm{d}m\right|\\
\le&\sup_{0\le r<t}\tilde{\varepsilon}^{-1}\int_r^{r+(t-s)\tilde{\varepsilon}} \left|\mathbb{E}_{V_{r/\varepsilon},x=\theta_r^{\varepsilon}}\Big[\big\langle\nabla_\theta\varphi(x),\nabla_\theta f(x,V_{(m-r)/\varepsilon})-g(x)\big\rangle\Big]\right|\mathrm{d}m\\
\le&\sup_{0\le r<t}\tilde{\varepsilon}^{-1}\left\|\langle\nabla_\theta\varphi(\theta_r^{\varepsilon}),\nabla_\theta f(\theta_r^{\varepsilon},\cdot)\rangle\right\|_\infty\int_r^{r+(t-s)\tilde{\varepsilon}} e^{-\delta(m-r)/\varepsilon}dm\\
=&\sup_{0\le r<t}\tilde{\varepsilon}^{-1}\left\|\langle\nabla_\theta\varphi(\theta_r^{\varepsilon}),\nabla_\theta f(\theta_r^{\varepsilon},\cdot)\rangle\right\|_\infty\int_0^{(t-s)\tilde{\varepsilon}} e^{-\delta k/\varepsilon}dk\\
\le&C_{t,\varphi,f}\frac{\varepsilon}{\delta\tilde{\varepsilon}}\le C_{t,\varphi,f}\frac{\sqrt{\varepsilon}}{2\delta}\to 0.
\end{aligned}
$$

This completes the proof of (13). Hence, any weak limit of $(\theta_t^{\varepsilon})_{t\ge0}$ is a martingale solution to equation (9). Since equation (9) is a deterministic ordinary differential equation and $\theta_0^{\varepsilon}=\theta_0$ is independent of ε, we have $(\theta_t^{\varepsilon})_{t\ge0}$ converges weakly to $(\zeta_t)_{t\ge0}$ as $\varepsilon\to0$. ■

In general, the constant $C_{t,\varphi,f}$ cannot be controlled since we only assume $\nabla_\theta f$ is Lipschitz in θ. In this case, θ_t is not bounded uniformly in time. We shall study the convergence rate under Wasserstein distance in Corollary 4, which is also not uniform in time, yet goes to 0 as $\varepsilon\to0$. In section 3.2, we shall show that with an additional assumption, this constant can be controlled and we obtain a longtime result as $t\to\infty$ with fixed ε. Moreover, the obtained upper bound of the Wasserstein distance between the invariant states goes to 0 as $\varepsilon\to0$. See Proposition 3 for more details.

Before stating Corollary 4, we need to introduce some notation. Let ν and ν' be two probability measures on $(\mathbb{R}^K,\mathcal{B}(\mathbb{R}^K))$. We define the Wasserstein distance between those measures by

$$\mathcal{W}_d(\nu,\nu')=\inf_{\Gamma\in\mathcal{H}(\nu,\nu')}\int_{\mathbb{R}^K\times\mathbb{R}^K} d(y,y')\Gamma(\mathrm{d}y,\mathrm{d}y'),$$

where $d(y,y'):=1\wedge\|y-y'\|$ and $\mathcal{H}(\nu,\nu')$ is the set of coupling between ν and ν', i.e.

$$\mathcal{H}(\nu,\nu')=\{\Gamma\in\Pr(\mathbb{R}^K\times\mathbb{R}^K):\Gamma(A\times\mathbb{R}^K)=\nu(A),\Gamma(\mathbb{R}^K\times B)=\nu'(B),\forall A,B\in\mathcal{B}(\mathbb{R}^K)\}.$$

To simplify the notation, for $B\in\mathcal{B}(\mathbb{R}^K)$, $\theta\in\mathbb{R}^K$, and $y\in S$, we denote

$$C_t^{\varepsilon}(B|\theta,y):=\mathbb{P}_y(\theta_t^{\varepsilon}\in B|\theta_0^{\varepsilon}=\theta),$$

$$C_t^\varepsilon(B|\theta,\pi) := \mathbb{P}_\pi(\theta_t^\varepsilon \in B|\theta_0^\varepsilon = \theta),$$

where π is the invariant measure of $(V_t)_{t\geq 0}$.

Now we study the approximation property of SGPC in the Wasserstein distance. Indeed, the following corollary follows immediately from Theorem 3.

Corollary 4 *There exists a function $\alpha : (0,1) \to [0,1]$, such that*

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|\theta_0,\pi), \delta(\cdot - \zeta_t)) \leq (\exp(t)\alpha(\varepsilon)) \wedge 1$$

and $\lim_{\varepsilon\to 0}\alpha(\varepsilon) = 0$.

Proof By Theorem 3, we have $(\theta_t^\varepsilon)_{t\geq 0} \Rightarrow (\zeta_t)_{t\geq 0}$. By Skorokhod's representation theorem, there exists a sequence $(\tilde{\theta}_t^\varepsilon)_{t\geq 0}$ such that

$$\begin{aligned}&(\tilde{\theta}_t^\varepsilon)_{t\geq 0} \overset{d}{=} (\theta_t^\varepsilon)_{t\geq 0} \text{ under } \mathbb{P}_\pi,\\ &\rho\Big((\tilde{\theta}_t^\varepsilon - \zeta_t)_{t\geq 0}, 0\Big) \to 0 \text{ almost surely in } \mathbb{P}_\pi.\end{aligned}$$

This implies

$$\mathbb{E}_\pi[F((\tilde{\theta}_t^\varepsilon - \zeta_t)_{t\geq 0})] \to F(0),$$

for any bounded continuous function F on $\mathcal{C}([0,\infty) : \mathbb{R}^K)$. By taking

$$F((\tilde{\theta}_t^\varepsilon - \zeta_t)_{t\geq 0})) = \sup_{t\geq 0}\exp(-t)\left(1 \wedge \sup_{0\leq s\leq t}\left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\right)$$

and

$$\alpha(\varepsilon) := \mathbb{E}_\pi\Big[\sup_{t\geq 0}\exp(-t)\left(1 \wedge \sup_{0\leq s\leq t}\left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\right)\Big] \to 0,$$

we have, for all $t \geq 0$,

$$\mathbb{E}_\pi\Big[1 \wedge \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\Big] \leq \exp(t)\alpha(\varepsilon).$$

Since $1 \wedge \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\| \leq 1$ and $\tilde{\theta}_t^\varepsilon \overset{d}{=} \theta_t^\varepsilon$, denoting the distribution of $\tilde{\theta}_t^\varepsilon$ as $F_{\tilde{\theta}_t^\varepsilon}$,

$$\begin{aligned}\mathcal{W}_d(C_t^\varepsilon(\cdot|\theta_0,\pi), \delta(\cdot - \zeta_t)) \leq& \mathcal{W}_d(F_{\tilde{\theta}_t^\varepsilon}, C_t^\varepsilon(\cdot|\theta_0,\pi)) + \mathbb{E}_\pi\Big[1 \wedge \left\|\tilde{\theta}_t^\varepsilon - \zeta_t\right\|\Big]\\ \leq& (\exp(t)\alpha(\varepsilon)) \wedge 1.\end{aligned}$$

■

Finally in this section, we look at a technical result concerning the asymptotic behavior of the full gradient flow $(\zeta_t)_{t\geq 0}$. First, we will additionally assume that the subsampled target function $f(\cdot, y)$ in the optimization problem is strongly convex, with a convexity parameter that does not depend on $y \in S$. We state this assumption below.

Assumption 3 (Strong Convexity) *For any $x_1, x_2 \in \mathbb{R}^K$,*

$$\langle x_1 - x_2, \nabla_\theta f(x_1, y) - \nabla_\theta f(x_2, y)\rangle \geq \kappa \|x_1 - x_2\|^2$$

where $\kappa > 0$ and κ is independent of $y \in S$.

Strong convexity implies, of course, that the full target function $g := \int_S \nabla_\theta f(\cdot, y)\pi(\mathrm{d}y)$ has a unique minimizer θ^*. It also implies that the associated full gradient flow $(\zeta_t)_{t\geq 0}$ converges at exponential speed to this unique minimizer. We give a short proof of this statement below.

Lemma 5 *Let $(\zeta_t)_{t\geq 0}$ be the process that solves (9) with initial data θ_0. Under Assumption 3, we have*

$$\|\zeta_t - \theta_*\|^2 \leq \|\theta_0 - \theta_*\|^2 \exp(-\kappa t),$$

where θ_* is a stationary solution of (9).

Proof Since θ_* is a stationary solution,

$$g(\theta_*) = 0 \quad \text{and} \quad \mathrm{d}(\zeta_t - \theta_*) = -(g(\zeta_t) - g(\theta_*))\mathrm{d}t.$$

Therefore,

$$\frac{\mathrm{d}\|\zeta_t - \theta_*\|^2}{\mathrm{d}t} = 2\left\langle \zeta_t - \theta_*, \frac{\mathrm{d}(\zeta_t - \theta_*)}{\mathrm{d}t}\right\rangle = -2\langle \zeta_t - \theta_*, g(\zeta_t) - g(\theta_*)\rangle \leq -2\kappa\|\zeta_t - \theta_*\|^2.$$

By Grönwall's inequality, we have $\|\zeta_t - \theta_*\|^2 \leq \|\theta_0 - \theta_*\|^2 \exp(-\kappa t)$. ■

3.2 Longtime behavior and ergodicity

We now study the longtime behavior of SGPC, i.e. the behavior and distribution of $(\theta^\varepsilon_t, V_{t/\varepsilon})$ for $t \gg 0$ large. Indeed, the main result of this section will be the geometric ergodicity of this coupled process and a study of its stationary measure. Initially, we study stability of the stochastic gradient process $(\theta^\varepsilon_t)_{t\geq 0}$.

Lemma 6 *Under Assumption 3, we have*

$$\|\theta^\varepsilon_t\|^2 \leq \|\theta^\varepsilon_0\|^2 \exp(-\kappa t) + \frac{8K_f^2}{\kappa^2},$$

where $K_f := \sup_{y\in S}\|\nabla_\theta f(0, y)\|$.

Proof By Itô's formula, we have

$$\frac{\mathrm{d}\|\theta^\varepsilon_t\|^2}{\mathrm{d}t} = 2\langle \theta^\varepsilon_t, d\theta^\varepsilon_t/dt\rangle = -2\left\langle \theta^\varepsilon_t, \nabla_\theta f(\theta^\varepsilon_t, V_{t/\varepsilon})\right\rangle. \tag{16}$$

Assumption 3 implies,

$$\begin{aligned}\left\langle \theta_t^\varepsilon, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon})\right\rangle =& \left\langle \theta_t^\varepsilon - 0, \nabla_\theta f(\theta_t^\varepsilon, V_{t/\varepsilon}) - \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle + \left\langle \theta_t^\varepsilon, \nabla_\theta f(0, V_{t/\varepsilon})\right\rangle \\ \geq& \kappa \left\|\theta_t^\varepsilon\right\|^2 - \left\|\theta_t^\varepsilon\right\| \left\|\nabla_\theta f(0, V_{t/\varepsilon})\right\| \\ \geq& \frac{\kappa}{2} \left\|\theta_t^\varepsilon\right\|^2 - \frac{4}{\kappa} \left\|\nabla_\theta f(0, V_{t/\varepsilon})\right\|^2 \\ \geq& \frac{\kappa}{2} \left\|\theta_t^\varepsilon\right\|^2 - \frac{4K_f^2}{\kappa}.\end{aligned}$$

Hence, (16) implies

$$\frac{\mathrm{d}\left\|\theta_t^\varepsilon\right\|^2}{\mathrm{d}t} \leq -\kappa \left\|\theta_t^\varepsilon\right\|^2 + \frac{8K_f^2}{\kappa}. \tag{17}$$

Multiplying $\exp(\kappa t)$ on both sides of (17), we get

$$\frac{\mathrm{d}(\left\|\theta_t^\varepsilon\right\|^2 \exp(\kappa t))}{\mathrm{d}t} \leq \frac{8K_f^2 \exp(\kappa t)}{\kappa},$$

that is

$$\left\|\theta_t^\varepsilon\right\|^2 \exp(\kappa t) - \left\|\theta_0^\varepsilon\right\|^2 \leq \frac{8K_f^2(\exp(\kappa t) - 1)}{\kappa^2} \leq \frac{8K_f^2 \exp(\kappa t)}{\kappa^2}$$

and therefore,

$$\left\|\theta_t^\varepsilon\right\|^2 \leq \left\|\theta_0^\varepsilon\right\|^2 \exp(-\kappa t) + \frac{8K_f^2}{\kappa^2}.$$

■

Using this lemma, we are now able to prove the first main result of this section, showing geometric ergodicity of $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$. First, we introduce a Wasserstein distance, on the space on which $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$ lives. Let Π and Π' be two probability measures on $(\mathbb{R}^K \times S, \mathcal{B}(\mathbb{R}^K \times S))$. We define the Wasserstein distance between those measures by

$$\widetilde{\mathcal{W}}_{\tilde{d}}(\Pi, \Pi') = \inf_{\widetilde{\Gamma} \in \mathcal{H}(\Pi, \Pi')} \int_{(\mathbb{R}^K \times S) \times (\mathbb{R}^K \times S)} \tilde{d}((u, v), (u', v')) \widetilde{\Gamma}(\mathrm{d}u\mathrm{d}v, \mathrm{d}u'\mathrm{d}v'),$$

where $\tilde{d}((u, v), (u', v')) := \mathbf{1}_{v \neq v'} + (1 \wedge \|u - u'\|)\mathbf{1}_{v = v'}$. Intuitively, when the indices v and v' are different, the distance should be large regardless of the distance between u and u'; when $v = v'$, the distance coincides with $d(\cdot, \cdot)$. For $a \in S$ and $m \in \mathbb{R}^K$, let $H_t^\varepsilon(\cdot | m, a)$ be the distribution of $(\theta_t^\varepsilon, V_{t/\varepsilon})$ under $\mathbb{P}_a$ with $\theta_0^\varepsilon = m$. Moreover, recall that $(V_t)_{t\geq 0}$ is a Feller process that satisfies Assumption 1. More specifically, it satisfies Assumption 1 (iii) with a constant δ:

$$\sup_{x \in S} \tilde{\mathbb{P}}(T^x \geq t) \leq C \exp(-\delta t),$$

where $T^x := \inf\{t \geq 0 \mid V_t^x = V_t^\pi\}$. With δ defined in this way, we have the following theorem.

Theorem 7 *Under Assumption 3, for any $0 < \varepsilon \leq 1 \wedge (\delta/2\kappa)$, the (coupled) process $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$ admits an unique stationary measure Π^ε on $(\mathbb{R}^K \times S, \mathcal{B}(\mathbb{R}^K \times S))$. Moreover,*

$$\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,a), \Pi^\varepsilon) \leq C_f \exp(-\kappa t) \int_{\mathbb{R}^K} (1 + \|x-m\|^2) \Pi^\varepsilon(\mathrm{d}x, S), \tag{18}$$

$$\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,\pi), \Pi^\varepsilon) \leq C_f \exp(-\kappa t) \int_{\mathbb{R}^K} \|x-m\|^2 \Pi^\varepsilon(\mathrm{d}x, S), \tag{19}$$

where the constant C_f only depends on f.

Proof To obtain the existence of the invariant measure, we apply the weak form of Harris' Theorem in Cloez and Hairer (2015, Theorem 3.7) by verifying the Lyapunov condition, the $\tilde{d}$-contracting condition, and the $\tilde{d}$-small condition.
(i) **Lyapunov condition:** Let $V(x,y) = \|x\|^2$. To verify that it satisfies (2.1) in Cloez and Hairer (2015, Definition 2.1), we take $x = \theta_0^\varepsilon$ and $(P_t)_{t\geq 0}$ to be the semi-group associated with $(\theta_t^\varepsilon)_{t\geq 0}$. Then Lemma 6 yields that $V(x,y)$ is a Lyapunov function. The existence of the Lyapunov function especially prevents the coupled processes from going to infinity.
(ii) **$\tilde{d}$-contracting condition:** $\tilde{d}$-contracting states that there exists $t^* > 0$ such that for any $t > t^*$, there exists some $\alpha < 1$ such that

$$\widetilde{\mathcal{W}}_{\tilde{d}}(\delta_{(m,a)} P_t^\varepsilon, \delta_{(n,b)} P_t^\varepsilon) \leq \alpha \tilde{d}((m,a),(n,b))$$

for any $(m,a),(n,b) \in \mathbb{R}^K \times S$ such that $\tilde{d}((m,a),(n,b)) < 1$. Here, $(P_t^\varepsilon)_{t\geq 0}$ is the semi-group operator associated with (8). Notice that $\tilde{d}((m,a),(n,b)) < 1$ implies $a = b$. Let $(\theta_t^{(m,a)})_{t\geq 0}$, $(\theta_t^{(n,a)})_{t\geq 0}$ solve the following equations:

$$\theta_t^{(m,a)} = m - \int_0^t \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{s/\varepsilon}^a) \mathrm{d}s,$$
$$\theta_t^{(n,a)} = n - \int_0^t \nabla_\theta f(\theta_t^{(n,a)}, \tilde{V}_{s/\varepsilon}^a) \mathrm{d}s,$$

where $\tilde{V}_t^a = V_t^a$ for $t \leq T^a$ and $\tilde{V}_t^a = V_t^\pi$ for $t > T^a$. Then by Itô's formula and Assumption 3, we obtain

$$\begin{aligned}\mathrm{d}\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2 / \mathrm{d}t &= -2\left\langle \theta_t^{(m,a)} - \theta_t^{(n,a)}, \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{s/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,a)}, \tilde{V}_{s/\varepsilon}^a)\right\rangle \\ &\leq -\kappa \left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2.\end{aligned}$$

By Grönwall's inequality, we have

$$\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2 \leq \exp(-\kappa t) \|m-n\|^2. \tag{20}$$

Noticing that $\tilde{d}((m,a),(n,b)) < 1$ implies $\|m-n\|^2 < 1$, by choosing $t \geq \frac{1}{\kappa}$, we obtain

$$\begin{aligned}\left\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\right\|^2 &\leq \exp(-1) \|m-n\|^2 \\ &= \exp(-1)(\|m-n\|^2 \wedge 1) = \exp(-1)\tilde{d}^2((m,a),(n,a)).\end{aligned}$$

Therefore, with $t^* = 1/\kappa$,

$$\widetilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,a), H_t^\varepsilon(\cdot|n,b)) \le \tilde{\mathbb{E}}\Big[\, \Big\|\theta_t^{(m,a)} - \theta_t^{(n,a)}\Big\| \,\Big] \le \exp(-1)\tilde{d}((m,a),(n,b)),$$

where $\tilde{\mathbb{E}}$ is the expectation under $\tilde{\mathbb{P}}$ (see Assumption 1).
(iii) $\tilde{d}$**-small condition:** We shall verify that there exists $t_* > 0$ such that for any $t > t_*$, the sublevel set $\mathscr{V} := \{(x,y) \in \mathbb{R}^K \times S \mid V(x,y) \le 32K_f^2/\kappa^2\}$ is $\tilde{d}$-small for $(P_t^\varepsilon)_{t\ge 0}$, meaning that there exists a constant ζ such that

$$\widetilde{\mathcal{W}}_{\tilde{d}}(\delta_{(m,a)}P_t^\varepsilon, \delta_{(n,b)}P_t^\varepsilon) \le 1 - \zeta,$$

for all $(m,a),(n,b) \in \mathscr{V}$. Let $(\theta_t^{(m,a)})_{t\ge 0},\ (\theta_t^{(n,b)})_{t\ge 0}$ solve the following equations:

$$\theta_t^{(m,a)} = m - \int_0^t \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{s/\varepsilon}^a)\mathrm{d}s,$$
$$\theta_t^{(n,b)} = n - \int_0^t \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{s/\varepsilon}^b)\mathrm{d}s,$$

where $\tilde{V}_t^a = V_t^a$ for $t \le T^a$ and $\tilde{V}_t^a = V_t^\pi$ for $t > T^a$; $\tilde{V}_t^b = V_t^b$ for $t \le T^b$ and $\tilde{V}_t^b = V_t^\pi$ for $t > T^b$. By Itô's formula, Assumption 3, and the ε-Young inequality,

$$\begin{aligned}
&\mathrm{d}\left\|\theta_t^{(m,a)} - \theta_t^{(n,b)}\right\|^2/\mathrm{d}t \\
=& -2\left\langle \theta_t^{(m,a)} - \theta_t^{(n,b)}, \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\rangle \\
=& -2\left\langle \theta_t^{(m,a)} - \theta_t^{(n,b)}, \nabla_\theta f(\theta_t^{(m,a)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^a)\right\rangle \\
&-2\left\langle \theta_t^{(m,a)} - \theta_t^{(n,b)}, \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\rangle \\
\le& -2\kappa\left\|\theta_t^{(m,a)} - \theta_t^{(n,b)}\right\|^2 + 2\left|\left\langle \theta_t^{(m,a)} - \theta_t^{(n,b)}, \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\rangle\right| \\
\le& -\kappa\left\|\theta_t^{(m,a)} - \theta_t^{(n,b)}\right\|^2 + \frac{4}{\kappa}\left\|\nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\|^2.
\end{aligned}$$

Multiplying $\exp(\kappa t)$ on both sides, we obtain

$$\mathrm{d}\Big(\exp(\kappa t)\left\|\theta_t^{(m,a)} - \theta_t^{(n,b)}\right\|^2\Big)/\mathrm{d}t \le \frac{4\exp(\kappa t)}{\kappa}\left\|\nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_t^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\|^2,$$

that is

$$\begin{aligned}
\exp(\kappa t)\left\|\theta_t^{(m,a)} - \theta_t^{(n,b)}\right\|^2 \le& \|m-n\|^2 \\
&+ \frac{4}{\kappa}\int_0^t \exp(\kappa s)\left\|\nabla_\theta f(\theta_s^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_s^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\|^2 \mathrm{d}s.
\end{aligned}$$

Notice that $\tilde{V}_t^a = \tilde{V}_t^b$ if $t > T^a \vee T^b$. Hence, we have

$$\begin{aligned}
&\exp(\kappa t)\left\|\theta_t^{(m,a)} - \theta_t^{(n,b)}\right\|^2 \\
&\quad\le \|m-n\|^2 + \frac{4}{\kappa}\int_0^{t\wedge\varepsilon(T^a\vee T^b)} \exp(\kappa s)\left\|\nabla_\theta f(\theta_s^{(n,b)}, \tilde{V}_{t/\varepsilon}^a) - \nabla_\theta f(\theta_s^{(n,b)}, \tilde{V}_{t/\varepsilon}^b)\right\|^2 \mathrm{d}s.
\end{aligned}$$

By Lemma 6, we have, for $(m,a),(n,b)\in\mathscr{V}\times S$,

$$\left\|\nabla_\theta f(\theta_t^{(m,a)},\tilde{V}_{t/\varepsilon}^a)\right\| \text{ and } \left\|\nabla_\theta f(\theta_t^{(n,b)},\tilde{V}_{t/\varepsilon}^b)\right\|$$

are bounded by some constant C_f. This implies

$$\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2 \le \exp(-\kappa t)\left\|m-n\right\|^2 + 4C_f\exp(-\kappa t)\exp(\kappa\varepsilon(T^a\vee T^b)). \tag{21}$$

Recall $0<\varepsilon\le 1\wedge\frac{\delta}{2\kappa}$. By (iii), Assumption 1,

$$\begin{aligned}\tilde{\mathbb{E}}[\exp(\kappa\varepsilon(T^a\vee T^b))] =& \kappa\varepsilon\int_0^\infty \exp(\kappa\varepsilon x)\tilde{\mathbb{P}}(T^a\vee T^b\ge x)\mathrm{d}x\\ \le& C\kappa\varepsilon\int_0^\infty \exp(\kappa\varepsilon x)\exp(-\delta x)\mathrm{d}x\\ \le& \frac{C\delta}{2}\int_0^\infty \exp\left(-\frac{\delta x}{2}\right)\mathrm{d}x = C.\end{aligned}$$

Therefore,

$$\tilde{\mathbb{E}}\Big[\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2\Big] \le \exp(-\kappa t)\left\|m-n\right\|^2 + C_f\exp(-\kappa t).$$

Moreover,

$$\tilde{\mathbb{P}}(\tilde{V}_{t/\varepsilon}^a\neq\tilde{V}_{t/\varepsilon}^b) = \tilde{\mathbb{P}}(T^a\vee T^b\ge t/\varepsilon)\le C\exp(-\delta t/\varepsilon)\le C\exp(-2\kappa t).$$

Hence for any $(m,a),(n,b)\in\mathscr{V}\times S$, by taking $t\ge\frac{1}{\kappa}[\log(8C_f)+\log(512K_f^2/\kappa^2)+\frac{1}{2}\log(8C)]$, we have

$$\widetilde{\mathcal{W}}_{\bar{d}}^2(H_t^\varepsilon(\cdot|m,a),H_t^\varepsilon(\cdot|n,b)) \le \tilde{\mathbb{P}}(\tilde{V}_{t/\varepsilon}^a\neq\tilde{V}_{t/\varepsilon}^b) + \tilde{\mathbb{E}}\Big[\left\|\theta_t^{(m,a)}-\theta_t^{(n,b)}\right\|^2\Big]\le\frac{1}{2}.$$

We have verified all three conditions from Cloez and Hairer (2015, Theorem 3.7) and hence conclude the existence and uniqueness of the invariant measure of $(\theta_t^\varepsilon,V_{t/\varepsilon})_{t\ge0}$ denoted as Π^ε.

Next, we are going to prove (18) and (19). Define Θ^ε such that $(\Theta^\varepsilon,\tilde{V}_0^\pi)\sim\Pi^\varepsilon$ in $\tilde{\mathbb{P}}$. Let $\theta_t^{\Pi^\varepsilon}$ and $\theta_t^{(m,\pi)}$ solve following equations

$$\begin{aligned}\theta_t^{\Pi^\varepsilon} &= \Theta^\varepsilon - \int_0^t \nabla_\theta f(\theta_t^{\Pi^\varepsilon},\tilde{V}_{t/\varepsilon}^\pi)\mathrm{d}s,\\ \theta_t^{(m,\pi)} &= m - \int_0^t \nabla_\theta f(\theta_t^{(m,\pi)},\tilde{V}_{t/\varepsilon}^\pi)\mathrm{d}s.\end{aligned}$$

Recall that from (21) and (20), we have

$$\begin{aligned}\left\|\theta_t^{(m,a)}-\theta_t^{\Pi^\varepsilon}\right\|^2 &\le \exp(-\kappa t)\left\|m-\Theta^\varepsilon\right\|^2 + 4C_f\exp(-\kappa t)\exp(\kappa\varepsilon T^a),\\ \left\|\theta_t^{(m,\pi)}-\theta_t^{\Pi^\varepsilon}\right\|^2 &\le \exp(-\kappa t)\left\|m-\Theta^\varepsilon\right\|^2.\end{aligned}$$

Therefore, by (iii), Assumption 1 and since $0 < \varepsilon \leq 1 \wedge (\delta/2\kappa)$, we have

$$\begin{aligned}
\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon) \leq& \tilde{\mathbb{P}}(\tilde{V}_{t/\varepsilon}^a \neq \tilde{V}_{t/\varepsilon}^\pi) + \tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,a)} - \theta_t^{\Pi^\varepsilon}\Big\|^2\Big] \\
\leq& \tilde{\mathbb{P}}(T^a \geq \frac{t}{\varepsilon}) + \exp(-\kappa t)\tilde{\mathbb{E}}[\|m - \Theta^\varepsilon\|^2] + 4C_f \exp(-\kappa t)\tilde{\mathbb{E}}[\exp(\kappa\varepsilon T^a)] \\
\leq& C_f \exp(-\kappa t) \int_{\mathbb{R}^K} (1 + \|x - m\|^2)\Pi^\varepsilon(\mathrm{d}x, S),
\end{aligned}$$

$$\begin{aligned}
\widetilde{\mathcal{W}}_{\tilde{d}}^2(H_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon) \leq& \tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,\pi)} - \theta_t^{\Pi^\varepsilon}\Big\|^2\Big] \\
\leq& \exp(-\kappa t)\tilde{\mathbb{E}}[\|m - \Theta^\varepsilon\|^2] \\
\leq& C_f \exp(-\kappa t) \int_{\mathbb{R}^K} \|x - m\|^2 \Pi^\varepsilon(\mathrm{d}x, S).
\end{aligned}$$

■

In the following corollaries, we study the integrals on the right-hand side of the inequalities in Theorem 7. Moreover, we show that the result above immediately implies not only geometric ergodicity of the coupled process $(\theta_t^\varepsilon, V_{t/\varepsilon})_{t\geq 0}$, but also of its marginal, the stochastic gradient process $(\theta_t^\varepsilon)_{t\geq 0}$.

Corollary 8 *Under the same assumptions as Theorem 7, there exists a constant $C_{f,m}$ that depends only on f and the initial value $m = \theta_0^\varepsilon$, such that*

$$\widetilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon) \leq C_{f,m} \exp\left(-\frac{\kappa t}{2}\right), \tag{22}$$

$$\widetilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon) \leq C_{f,m} \exp\left(-\frac{\kappa t}{2}\right), \tag{23}$$

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon(\cdot,S)) \leq C_{f,m} \exp\left(-\frac{\kappa t}{2}\right), \tag{24}$$

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon(\cdot,S)) \leq C_{f,m} \exp\left(-\frac{\kappa t}{2}\right). \tag{25}$$

Proof By Lemma 6,

$$\int_{\|x\|^2 \geq \frac{8K^2}{\kappa^2} + \|m\|^2 + 1} \Pi^\varepsilon(\mathrm{d}x, S) = \lim_{t\to\infty} \tilde{\mathbb{P}}(\Big\|\theta_t^{(m,a)}\Big\| \geq \frac{8K_f^2}{\kappa^2} + \|m\|^2 + 1) = 0.$$

Let $C_{f,m} = C_f(2\|m\| + \frac{4K_f}{\kappa} + 2)$. From (18), we have

$$\begin{aligned}
\tilde{\mathcal{W}}_{\tilde{d}}(H_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon) \leq& C_f \exp\left(\frac{-\kappa t}{2}\right)\Big(\int_{\mathbb{R}^K} (1 + \|x - m\|^2)\Pi^\varepsilon(\mathrm{d}x, S)\Big)^{1/2} \\
=& C_f \exp\left(\frac{-\kappa t}{2}\right)\Big(\int_{\|x\|^2 \leq \frac{8K_f^2}{\kappa^2} + \|m\|^2 + 1} (1 + \|x - m\|^2)\Pi^\varepsilon(\mathrm{d}x, S)\Big)^{1/2} \\
\leq& C_{f,m} \exp\left(\frac{-\kappa t}{2}\right).
\end{aligned}$$

(23) can be derived similarly from (19). Moreover, notice that

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,a),\Pi^\varepsilon(\cdot,S)) \leq \tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,a)} - \theta_t^{\Pi^\varepsilon}\Big\|\Big], \tag{26}$$

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi),\Pi^\varepsilon(\cdot,S)) \leq \tilde{\mathbb{E}}\Big[\Big\|\theta_t^{(m,\pi)} - \theta_t^{\Pi^\varepsilon}\Big\|\Big]. \tag{27}$$

(24) and (25) can be derived similarly from (26) and (27). ■

Combining (24) and (25), we immediately have:

Corollary 9 *Under the same assumptions as Theorem 7,*

$$\mathcal{W}_d(C_t^\varepsilon(\cdot|m,a),C_t^\varepsilon(\cdot|m,\pi)) \leq C_{f,m}\exp\left(-\frac{\kappa t}{2}\right).$$

So far, we have shown that the stochastic gradient process $(\theta_t^\varepsilon)_{t\geq 0}$ converges to a unique stationary measure $\Pi^\varepsilon(\cdot,S)$. It is often not possible to determine this stationary measure. However, we can comment on its asymptotic behavior as $\varepsilon \to 0$. Indeed, we will show that $\Pi^\varepsilon(\cdot,S)$ concentrates around the minimizer θ_* of the full target function.

Proposition 3 *Under Assumption 3, the measure $\Pi^\varepsilon(\cdot,S)$ on $(\mathbb{R}^K,\mathcal{B}(\mathbb{R}^K))$ approximates $\delta(\cdot-\theta_*)$. In other words, we have*

$$\mathcal{W}_d(\Pi^\varepsilon(\cdot,S),\delta(\cdot-\theta_*)) \leq \rho(\varepsilon)$$

where $\rho:(0,1)\to[0,1]$ and $\lim_{\varepsilon\to 0}\rho(\varepsilon)=0$.

Proof By the triangle inequality,

$$\begin{aligned}&\mathcal{W}_d(\Pi^\varepsilon(\cdot,S),\delta(\cdot-\theta_*))\\ \leq&\mathcal{W}_d(\Pi^\varepsilon(\cdot,S),C_t^\varepsilon(\cdot|m,\pi)) + \mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi),\delta(\cdot-\zeta_t)) + \mathcal{W}_d(\delta(\cdot-\zeta_t),\delta(\cdot-\theta_*)).\end{aligned}$$

Let $\theta_0 = \theta_0^\varepsilon = \theta_*$. Then by Lemma 5, we have the last term

$$\mathcal{W}_d(\delta(\cdot-\zeta_t),\delta(\cdot-\theta_*)) \leq \|\theta_*-\theta_*\|\exp(-\kappa t) = 0.$$

By (22) and Corollary 4, for any $t\geq 0$,

$$\mathcal{W}_d(\Pi^\varepsilon(\cdot,S),C_t^\varepsilon(\cdot|m,\pi)) + \mathcal{W}_d(C_t^\varepsilon(\cdot|m,\pi),\delta(\cdot-\zeta_t)) \leq C_{f,\theta_*}\exp\left(-\frac{\kappa t}{2}\right) + (\exp(t)\alpha(\varepsilon))\wedge 1.$$

By choosing $t = -\log(1\wedge\alpha(\varepsilon))/2$, we get

$$\begin{aligned}\mathcal{W}_d(\Pi^\varepsilon(\cdot,S),\delta(\cdot-\theta_*)) \leq& C_{f,\theta_*}\exp\left(-\frac{\kappa t}{2}\right) + (\exp(t)\alpha(\varepsilon))\wedge 1\\ \leq& C_{f,\theta_*}(\alpha(\varepsilon))^{\frac{\kappa}{4}} + (\alpha(\varepsilon))^{1/2}.\end{aligned}$$

Taking $\rho(\varepsilon) := \left(C_{f,\theta_*}(\alpha(\varepsilon))^{\frac{\kappa}{4}} + (\alpha(\varepsilon))^{1/2}\right)\wedge 1$ completes the proof. ■

4. Stochastic gradient processes with decreasing learning rate

Constant learning rates are popular in some practical situations, but the associated stochastic gradient process usually does not converge to the minimizer of Φ. This is also true for the discrete-time stochastic gradient descent algorithm. However, SGD can converge to the minimizer if the learning rate is decreased over time. In the following, we discuss a decreasing learning version of the stochastic gradient process and show that this dynamical system indeed converges to the minimizer of Φ.

As discussed in Section 1.2, we obtain the stochastic gradient process with decreasing learning rate by non-linearly rescaling the time in the constant-learning-rate index process $(V_t)_{t\geq 0}$. Indeed, we choose a function $\beta : [0,\infty) \to [0,\infty)$ and then define the decreasing learning rate index process by $(V_{\beta(t)})_{t\geq 0}$. We have discussed an intuitive way to construct a rescaling function β also in Section 1.2.

In the following, we define β through an integral $\beta(t) = \int_0^t \mu(s)\mathrm{d}s$, $t \geq 0$. We commence this section with necessary growth conditions on μ which allow us to then give the formal definition of the stochastic gradient process with decreasing learning rate. Then, we study the longtime behavior of this process.

Assumption 4 *Let $\mu : [0,\infty) \to (0,\infty)$ be a non-decreasing continuously differentiable function with $\lim_{t\to\infty} \mu(t) = \infty$ and*

$$\lim_{t\to\infty} \frac{\mu'(t)t}{\mu(t)} = 0.$$

Assumption 4 implies that μ goes to infinity, but at a very slow pace. Indeed it says that $\lim_{t\to\infty} \mu(t)/t^\gamma = 0$, $\gamma > 0$, that is μ grows slower than any polynomial.

Definition 2 *The* stochastic gradient process with decreasing learning rate (SGPD) *is a solution of the following stochastic differential equation,*

$$\begin{cases} \mathrm{d}\xi_t = -\nabla_\xi f(\xi_t, V_{\beta(t)})\mathrm{d}t, \\ \xi_0 = \theta_0, \end{cases} \tag{28}$$

where f satisfies Assumption 2, $(V_t)_{t\geq 0}$ is a Feller process that satisfies Assumption 1, and $\beta(t) = \int_0^t \mu(s)\mathrm{d}s$ with μ satisfying Assumption 4.

To see that $(\xi_t)_{t\geq 0}$ is well-defined, consider the following: $t \mapsto \beta(t)$ is an increasing continuous function. Thus, $(V_{\beta(t)})_{t\geq 0}$ is càdlàg and Feller with respect to $(\mathcal{F}_{\beta(t)})_{t\geq 0}$. We then obtain well-definedness of $(\xi_t)_{t\geq 0}$ by replacing $(V_{t/\varepsilon})_{t\geq 0}$ by $(V_{\beta(t)})_{t\geq 0}$ in the proof of Proposition 2.

We now move on to studying the longtime behavior of the SGPD $(\xi_t)_{t\geq 0}$. In a first technical result, we establish a connection between SGPD $(\xi_t)_{t\geq 0}$ and a time-rescaled version of SGPC $(\theta^\varepsilon_t)_{t\geq 0}$. To this end, note that $\dot\beta(t) = \mu(t) > 0$, $\beta(t)$ is strictly increasing. Hence, the inverse of $\beta(t)$ exists and

$$\beta^{-1}(t) = \int_0^t \frac{1}{\mu(\beta^{-1}(s))}\mathrm{d}s.$$

This gives us the following inequality.

Proposition 4 *For any $0 < \varepsilon < 1$,*

$$\left\|\xi_t - \theta^{\varepsilon}_{\varepsilon\beta(t)}\right\|^2 \le C_{f,\theta_0,\mu}\left[\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon} + \frac{1}{\varepsilon}\Big(\left|\frac{1}{\mu(t)} - \varepsilon\right| + \left|\frac{1}{\mu(\frac{t}{2})} - \varepsilon\right|\Big)\right]$$

almost surely, where the constant $C_{f,\theta_0,\mu}$ depends only on f, the initial data θ_0, and μ.

Proof From (28) and (8), we have

$$\begin{aligned}\xi_t =&\theta_0 - \int_0^{\beta(t)} \nabla_\xi f(\xi_{\beta^{-1}(s)}, V_s)\mathrm{d}\beta^{-1}(s),\\ \theta^{\varepsilon}_t =&\theta_0 - \varepsilon\int_0^{\frac{t}{\varepsilon}} \nabla_\theta f(\theta^{\varepsilon}_{\varepsilon s}, V_s)\mathrm{d}s.\end{aligned}$$

Let $b_t := \mathrm{d}\beta^{-1}(t)/\mathrm{d}t = 1/\mu(\beta^{-1}(t)) > 0$, we have

$$\begin{aligned}\xi_{\beta^{-1}(t)} =&\theta_0 - \int_0^{t} \nabla_\xi f(\xi_{\beta^{-1}(s)}, V_s)b_s\mathrm{d}s\\ \theta^{\varepsilon}_{\varepsilon t} =&\theta_0 - \varepsilon\int_0^{t} \nabla_\theta f(\theta^{\varepsilon}_{\varepsilon s}, V_s)\mathrm{d}s\end{aligned}$$

Therefore, by Itô's formula and Assumption 3,

$$\begin{aligned}\mathrm{d}\left\|\theta^{\varepsilon}_{\varepsilon t} - \xi_{\beta^{-1}(t)}\right\|^2/\mathrm{d}t =& -2\left\langle\theta^{\varepsilon}_{\varepsilon t} - \xi_{\beta^{-1}(t)}, \varepsilon\nabla_\theta f(\theta^{\varepsilon}_{\varepsilon t}, V_t) - \varepsilon\nabla_\xi f(\xi_{\beta^{-1}(t)}, V_t)\right\rangle\\ &-2(\varepsilon - b_t)\left\langle\theta^{\varepsilon}_{\varepsilon t} - \xi_{\beta^{-1}(t)}, \nabla_\xi f(\xi_{\beta^{-1}(t)}, V_t)\right\rangle\\ \le& -2\varepsilon\kappa\left\|\theta^{\varepsilon}_{\varepsilon t} - \xi_{\beta^{-1}(t)}\right\|^2 + C_{f,\theta_0}\left|b_t - \varepsilon\right|,\end{aligned}$$

where the last step follows from the boundedness of $\theta^{\varepsilon}_{\varepsilon t}$, $\xi_{\beta^{-1}(t)}$, and $\nabla_\xi f(\xi_{\beta^{-1}(t)}, V_t)$. $\xi_{\beta^{-1}(t)}$ is bounded can be showed similarly to Lemma 6. Multiplying $\exp(2\varepsilon\kappa t)$ on both sides, we obtain

$$\mathrm{d}\Big(\exp(2\varepsilon\kappa t)\left\|\theta^{\varepsilon}_{\varepsilon t} - \xi_{\beta^{-1}(t)}\right\|^2\Big)/\mathrm{d}t \le C_{f,\theta_0}\left|b_t - \varepsilon\right|\exp(2\varepsilon\kappa t),$$

which implies

$$\left\|\theta^{\varepsilon}_{\varepsilon t} - \xi_{\beta^{-1}(t)}\right\|^2 \le C_{f,\theta_0}\exp(-2\varepsilon\kappa t)\int_0^t \left|b_s - \varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s.$$

Notice that b_s is bounded and non-increasing, hence we have

$$\begin{aligned}\left\|\xi_t - \theta^{\varepsilon}_{\varepsilon\beta(t)}\right\|^2 \le& C_{f,\theta_0}\exp(-2\varepsilon\kappa\beta(t))\int_0^{\beta(t)}\left|b_s - \varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s\\ =&C_{f,\theta_0}\exp(-2\varepsilon\kappa\beta(t))\Big(\int_0^{\beta(\frac{t}{2})} + \int_{\beta(\frac{t}{2})}^{\beta(t)}\Big)\left|b_s - \varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s\\ \le&C_{f,\theta_0,\mu}\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon}\\ &+ C_{f,\theta_0}\exp(-2\varepsilon\kappa\beta(t))\int_{\beta(\frac{t}{2})}^{\beta(t)}\left|b_s - \varepsilon\right|\exp(2\varepsilon\kappa s)\mathrm{d}s\\ \le&C_{f,\theta_0,\mu}\left[\frac{\exp(-2\varepsilon\kappa(\beta(t)-\beta(\frac{t}{2})))}{\varepsilon} + \frac{1}{\varepsilon}\Big(\left|\frac{1}{\mu(t)} - \varepsilon\right| + \left|\frac{1}{\mu(\frac{t}{2})} - \varepsilon\right|\Big)\right].\end{aligned}$$

■

Now, we get to the main result of this section, where we show the convergence of $(\xi_t)_{t\geq 0}$ to the minimizer θ_* of Φ. In the following, we denote

$$D_t(B|\theta_0, a) := \mathbb{P}_a(\xi_t \in B|\xi_0 = \theta_0),$$
$$D_t(B|\theta_0, \pi) := \mathbb{P}_\pi(\xi_t \in B|\xi_0 = \theta_0) \qquad (B \in \mathcal{B}(\mathbb{R}^K), \theta_0 \in \mathbb{R}^K),$$

where $a \in S$ and π is the invariant measure of $(V_t)_{t\geq 0}$, respectively.

Theorem 10 *Under Assumption 3, given $\theta_0 \in \mathbb{R}^K$ and $a \in S$, there exists $T > 0$ such that for any $t > T$,*

$$\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), \delta(\cdot - \theta_*)) \leq C_{f,\theta_0,\mu} A(t), \tag{29}$$

$$\mathcal{W}_d(D_t(\cdot|\theta_0, a), \delta(\cdot - \theta_*)) \leq C_{f,\theta_0,\mu} A(t), \tag{30}$$

where

$$A(t) := \exp\left(\frac{-\kappa t}{8}\right) + \left[\frac{\mu(t) - \mu(\frac{t}{2})}{\mu(t)}\right]^{1/2} + \rho\left(\frac{1}{\mu(\frac{t}{2})}\right)$$

and $\lim_{t\to\infty} A(t) = 0$.

Proof To prove (29), by the triangle inequality,

$$\begin{aligned}&\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), \delta(\cdot - \theta_*))\\ \leq&\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi)) + \mathcal{W}_d(C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi), \Pi^\varepsilon(\cdot, S)) + \mathcal{W}_d(\Pi^\varepsilon(\cdot, S), \delta(\cdot - \theta_*)).\end{aligned}$$

For the last two terms, by (25) and Proposition 3,

$$\mathcal{W}_d(C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi), \Pi^\varepsilon(\cdot, S)) + \mathcal{W}_d(\Pi^\varepsilon(\cdot, S), \delta(\cdot - \theta_*)) \leq C_{f,m} \exp(-\kappa\varepsilon\beta(t)/2) + \rho(\varepsilon).$$

For the first term, by Proposition 4,

$$\begin{aligned}&\mathcal{W}_d(D_t(\cdot|\theta_0, \pi), C^\varepsilon_{\varepsilon\beta(t)}(\cdot|\theta_0, \pi))\\ &\qquad\leq C_{f,\theta_0,\mu}\left[\frac{\exp(-2\varepsilon\kappa(\beta(t) - \beta(\frac{t}{2})))}{\varepsilon} + \frac{1}{\varepsilon}\Big(\left|\frac{1}{\mu(t)} - \varepsilon\right| + \left|\frac{1}{\mu(\frac{t}{2})} - \varepsilon\right|\Big)\right]^{1/2}.\end{aligned}$$

Since $\lim_{t\to\infty} \mu(t) = \infty$, there exists $T > 0$ such that $1/\mu(\frac{T}{2}) < \frac{\delta}{2\kappa}$. Let $\varepsilon = 1/\mu(\frac{t}{2})$, $t > T$, we have

$$\exp\left(\frac{-\kappa\varepsilon\beta(t)}{2}\right) = \exp\left(\frac{-\kappa\int_0^t \mu(s)\mathrm{d}s}{2\mu(\frac{t}{2})}\right) \leq \exp\left(\frac{-\kappa t}{8}\right)$$

and

$$\begin{aligned}\frac{\exp\left(-2\varepsilon\kappa(\beta(t) - \beta(\frac{t}{2}))\right)}{\varepsilon} &= \mu\left(\frac{t}{2}\right)\exp\left(\frac{-\kappa\int_{\frac{t}{2}}^t \mu(s)\mathrm{d}s}{\mu(\frac{t}{2})}\right)\\ &\leq \mu\left(\frac{t}{2}\right)\exp\left(\frac{-\kappa t}{2}\right) \leq C\exp\left(\frac{-\kappa t}{8}\right).\end{aligned}$$

Therefore,

$$\mathcal{W}_d(D_t(\cdot|\theta_0,\pi),\delta(\cdot-\theta_*)) \le C_{f,\theta_0,\mu}\Big[\exp\left(\frac{-\kappa t}{8}\right)+\left(\frac{\mu(t)-\mu(\frac{t}{2})}{\mu(t)}\right)^{1/2}+\rho\left(\frac{1}{\mu\left(\frac{t}{2}\right)}\right)\Big]$$

From Assumption 4, by the mean value theorem,

$$\frac{\mu(t)-\mu(\frac{t}{2})}{\mu(t)} = \frac{t\mu'(\tau_t)}{2\mu(t)} = \frac{\tau_t\mu'(\tau_t)}{\mu(\tau_t)}\frac{t}{2\tau_t}\frac{\mu(\tau_t)}{\mu(t)} \le \frac{\tau_t\mu'(\tau_t)}{\mu(\tau_t)} \to 0$$

where $\tau_t \in [\frac{t}{2}, t]$. Thus, (29) is obtained by taking

$$A(t) := \exp\left(\frac{-\kappa t}{8}\right)+\left[\frac{\mu(t)-\mu(\frac{t}{2})}{\mu(t)}\right]^{1/2}+\rho\left(\frac{1}{\mu(\frac{t}{2})}\right).$$

To prove (30), we have

$$\begin{aligned}\mathcal{W}_d(D_t(\cdot|\theta_0,a),\delta(\cdot-\theta_*)) \le\ & \mathcal{W}_d(D_t(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a))\\ &+\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi))\\ &+\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi),\Pi^{\varepsilon}(\cdot,S))\\ &+\mathcal{W}_d(\Pi^{\varepsilon}(\cdot,S),\delta(\cdot-\theta_*)),\end{aligned}$$

by the triangle inequality. By Corollary 9, we have

$$\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi)) \le C_{f,m}\exp\left(-\frac{\kappa\varepsilon\beta(t)}{2}\right).$$

Notice that $C_{f,m}\exp(-\frac{\kappa t}{4}) \le C_{f,m}A(t)$ when $\varepsilon = 1/\mu(\frac{t}{2})$. Similar to the proof of (29), we have

$$\begin{aligned}\mathcal{W}_d(D_t(\cdot|\theta_0,a),C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,a))+\mathcal{W}_d(C^{\varepsilon}_{\varepsilon\beta(t)}(\cdot|\theta_0,\pi),\Pi^{\varepsilon}(\cdot,S))+\mathcal{W}_d(\Pi^{\varepsilon}(\cdot,S),\delta(\cdot-\theta_*))&\\ \le C_{f,\theta_0,\mu}A(t),&\end{aligned}$$

which completes the proof. ■

Thus, we have shown that the distribution of $(\xi_t)_{t\ge0}$ converges in Wasserstein distance to the Dirac measure concentrated in the minimizer θ_* of Φ. This result is independent of whether we initialize the index process $(V_{\beta(t)})_{t\ge0}$ with its stationary measure or with any deterministic value.

5. From continuous dynamics to practical optimization.

So far, we have discussed the stochastic gradient process as a continuous-time coupling of an ODE and a stochastic process. In order to apply the stochastic gradient process in practice, we need to discretize ODE and stochastic process with appropriate time-stepping schemes. That means, for a given increasing sequence $(t(k))_{k=0}^{\infty}$, with $t(0) := 0$

and $\lim_{k\to\infty} t(k) = \infty$, we seek discrete-time stochastic processes $(\widehat{V}_k, \widehat{\theta}_k)_{k=0}^{\infty}$, such that $(\widehat{V}_k, \widehat{\theta}_k)_{k=0}^{\infty} \approx (V_{t(k)}, \theta^{\varepsilon}_{t(k)})_{k=0}^{\infty}$ and analogous discretizations for $(V_{\beta(t)}, \xi_t)_{t\geq 0}$.

In the following, we propose and discuss time stepping strategies and the algorithms arising from them. We discuss the index process and gradient flow separately, which we consider sufficient as the coupling is only one-sided.

5.1 Discretization of the index process

We have defined the stochastic gradient process for a huge range of potential index processes $(V_t)_{t\geq 0}$. The discretization of such processes has been the topic of several works, see, e.g., Gillespie (1977); Lord et al. (2014). In the following, we focus on one case and refer to those previous works for other settings and details.

Indeed, we study the setting $S := [-1, 1]$ and $\pi := \mathrm{Unif}[-1, 1]$ and discuss the discretization of $(V_t)_{t\geq 0}$ as a Markov pure jump process and as a reflected Brownian motion.

Markov pure jump process.

A suitable Markov pure jump process is a piecewise constant càdlàg process $(V_t)_{t\geq 0}$ with Markov transition kernel

$$\mathbb{P}_x(V_t \in \cdot) = \exp(-\lambda t)\delta(\cdot - x) + (1 - \exp(-\lambda t))\mathrm{Unif}[-1, 1] \qquad (t \geq 0),$$

where $\lambda > 0$ is a rate parameter. We can now discretize the process $(V_t)_{t\geq 0}$ just through sampling from this Markov kernel for our discrete time points. We describe this in Algorithm 1.

Algorithm 1 Discretized Markov pure jump process

initialize $\widehat{V}_0$, $\lambda > 0$, and a sequence of points $(t(k))_{k=0}^{\infty}$
for $k = 1, 2, \ldots$ **do**
  sample $U \sim \mathrm{Unif}[0, 1]$
  if $U \leq \exp(-\lambda(t(k) - t(k-1)))$ **then**
    $\widehat{V}_k \leftarrow \widehat{V}_{k-1}$ {process stays at its current position}
  else
    sample $\widehat{V}_k \sim \mathrm{Unif}[-1, 1]$ {process jumps to a new position}
  end if
end for
return $(\widehat{V}_k)_{k=0}^{\infty}$

Reflected Brownian motion.

We have defined the reflected Brownian Motion on a non-empty compact interval through the Skorohod problem in Subsection 2.1.

Let $\sigma > 0$ and $(W_t)_{t\geq 0}$ be a standard Brownian motion. Probably the easiest way to sample a reflected Brownian motion is by discretizing the rescaled Brownian motion $(\sigma \cdot W_t)_{t\geq 0}$ using the Euler–Maruyama scheme and projecting back to S, whenever the

sequence leaves S. This scheme has been studied by Pettersson (1995). We describe the full scheme in Algorithm 2.

Pettersson (1995) shows that this scheme converges at a rather slow rate. As we usually assume that the domain on which we move is rather low-dimensional and the sampling is rather cheap, we can afford small discretization stepsizes $t(k) - t(k-1)$, for $k \in \mathbb{N}$. Thus, the slow rate of convergence is manageable. Other schemes for the discretization of reflected Brownian motions have been discussed by, e.g., Blanchet and Murthy (2018); Liu (1995).

Algorithm 2 Discretized Reflected Brownian motion on S

initialize $\widehat{V}_0$, $\sigma > 0$, a sequence of points $(t(k))_{k=0}^{\infty}$, and the projection operator proj_S mapping onto S
for $k = 1, 2, \ldots$ **do**
  $V' \leftarrow V_{k-1} + \sigma\sqrt{t(k) - t(k-1)}\psi, \quad \psi \sim \mathrm{N}(0, 1^2)$ {Euler-Maruyama update}
  if $V' \notin S$ **then**
    $\widehat{V}_k \leftarrow \mathrm{proj}_S V'$ {project back}
  else
    $\widehat{V}_k \leftarrow V'$ {accept Euler–Maruyama update}
  end if
end for
return $(\widehat{V}_k)_{k=0}^{\infty}$

5.2 Discretization of the gradient flow

We now briefly discuss the discretization of the gradient flow in the stochastic gradient process. Based on these ideas, we will conduct numerical experiments in Section 6.

Stochastic gradient descent

In stochastic gradient descent, the gradient flow is discretized with a forward Euler method. This method leads to an accurate discretization of the respective gradient flow if the stepsize/learning rates are sufficiently small. In the presence of rather large stepsizes and stiff vector fields, however, the forward Euler method may be inaccurate and unstable, see, e.g., Quarteroni et al. (2007).

Stability

Several ideas have been proposed to mitigate this problem. The stochastic proximal point method, for instance, uses the backward Euler method to discretize the gradient flow; see Bianchi (2015). Unfortunately, such implicit ODE integrators require us to invert a possibly highly non-linear and complex vector field. In convex stochastic optimization this inversion can be replaced by evaluating a proximal operator. For strongly convex optimization, on the other hand, Eftekhari et al. (2021) proposes stable explicit methods.

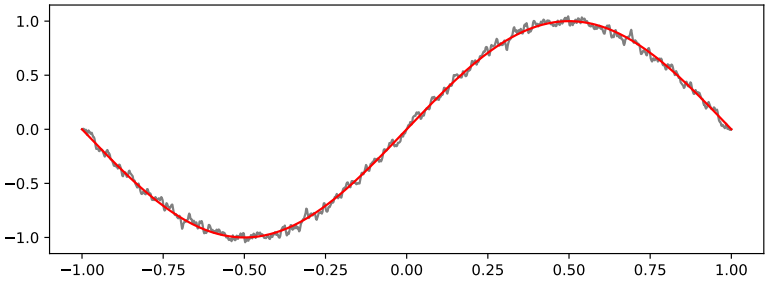

Figure 3: True function Θ (red) and noisy observation g (grey) in the polynomial regression example.

Efficient optimizers

Plenty of highly efficient methods for stochastic optimization methods are nowadays available, especially in machine learning. Those have often been proposed without necessarily thinking of the stable and accurate discretization of a gradient flow: such are adaptive methods (Kingma and Ba, 2015), variance reduced methods (e.g., Defazio et al., 2014), or momentum methods (e.g., Kovachki and Stuart, 2021 for an overview), which have been shown in multiple works to be highly efficient; partially also in non-convex optimization. We could understand those methods also as certain discretizations of the gradient flow. Thus, we may also consider the combination of a feasible index process $(V_t)_{t\geq 0}$ with the discrete dynamical system in, e.g., the Adam method (Kingma and Ba, 2015).

6. Applications

We now study two fields of application of the stochastic gradient process for continous data. In the first example, we consider regularized polynomial regression with noisy functional data. In this case, we can easily show that the necessary assumptions for our analysis hold. Thus, we use it to illustrate our analytical results and especially to learn about the implicit regularization that is put in place due to different index proccesses.

In the second example, we study so-called physics-informed neural networks. In these continuous-data machine learning problems, a deep neural network is used to approximate the solution of a partial differential equation. The associated optimization problem is usually non-convex. Our analysis does not hold in this case: We study it to get more insights in the behavior of the stochastic gradient process in state-of-the-art deep learning problems.

6.1 Polynomial regression with functional data

We begin with a simple polynomial regression problem with noisy functional data. We observe the function $g : [-1, 1] \to \mathbb{R}$ which is given through

$$g(y) := \Theta(y) + \Xi(y) \qquad (y \in [-1, 1]),$$

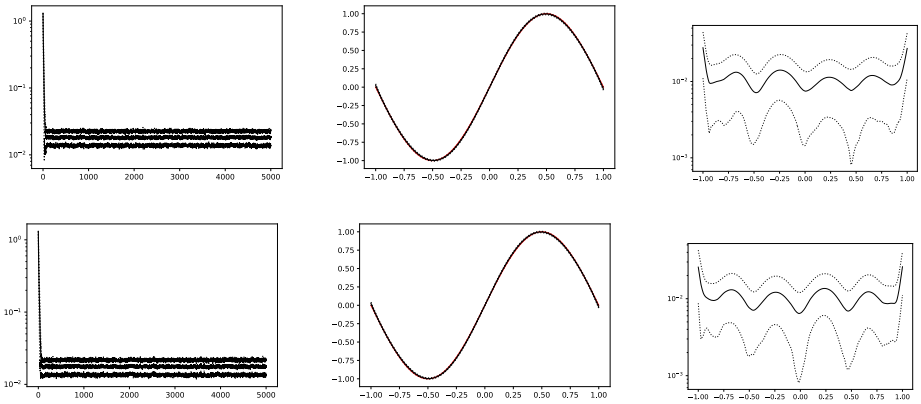

Figure 4: Estimation results of the polynomial regression problem using stochastic gradient descent with constant learning rate $\eta = 0.1$ (top row) and a version of stochastic gradient descent that uses the implicit midpoint rule (bottom row). The figures depict the mean over 100 runs (black solid line), mean $\pm$ standard deviation (black dotted line). Left column: trajectory of the rel_err over time; centre column: comparison of Θ (solid red line) and estimated polynomial; right column: estimation error in terms of abs_err.

where $\Theta : [-1, 1] \to \mathbb{R}$ is a smooth function and Ξ is a Gaussian process with highly oscillating, continuous realizations. We aim at identifying the unknown function Θ subject to the observational noise Ξ. Here, we represent the function Θ on a basis consisting of a finite number of Legendre polynomials on $[-1, 1]$. We denote this basis of Legendre polynomials by $(\ell_k)_{k=1}^K$. To estimate the prefactors of the polynomials, we minimize the potential

$$\Phi(\theta) := \frac{1}{2} \int_{[-1,1]} \left(g(y) - \sum_{k=1}^{K} \theta_k \ell_k(y) \right)^2 \mathrm{d}y + \frac{\alpha}{2} \|\theta\|_2^2 \qquad (\theta \in X), \tag{31}$$

where $\alpha > 0$ is a regularization parameter. This can be understood as a maximum-a-posteriori estimation of the unknown θ with Gaussian prior under the (misspecified) assumption that the data is perturbed with Gaussian white noise. We employ the following associated subsampled potentials:

$$f(\theta, y) := \frac{1}{2} \left(g(y) - \sum_{k=1}^{K} \theta_k \ell_k(y) \right)^2 + \frac{\alpha}{2} \|\theta\|_2^2 \qquad (\theta \in X, y \in [-1, 1]). \tag{32}$$

Those subsampled potentials satisfy the strong convexity assumption, i.e., Assumption 3.

Setup

In particular, we have produced artificial data g, by setting $\Theta := \sin(\pi\cdot)$ and choosing

$$\Xi(x) = \sum_{j=1}^{200} \frac{10}{1000 + (\pi j)^{3/2}} \sin(2\pi j(x - 0.5))\Xi_j \qquad (x \in [-1, 1])$$

and i.i.d. random variables $\Xi_1, \dots, \Xi_{200} \sim \mathrm{N}(0, 1^2)$. Note that Ξ is a Gaussian random field given through the truncated Karhunen-Loève expansion of a covariance operator that is related to the Matérn family, see, e.g., Lindgren et al. (2011).

We show Θ and g in Figure 3. For our estimation, we set $\alpha := 10^{-4}$ and use the $K = 9$ Legendre polynomials with degrees $0, \dots, 8$. We employ the stochastic gradient process with constant learning rate, using either a reflected diffusion process or a pure Markov jump process for the index process $(V_t)_{t\geq 0}$. We discretize the gradient flow using the implicit midpoint rule: an ODE $z' = q(z), z(0) = z_0$ is then discretized with stepsize $h > 0$ by successively solving the implicit formula

$$z_k = z_{k-1} + \frac{h}{2} q(z_k) + \frac{h}{2} q(z_{k-1}) \qquad (k \in \mathbb{N}).$$

In our experiments, we choose $h = 0.1$. We use Algorithms 1 and 2 to discretize the index processes with constant stepsize $t(\cdot) - t(\cdot - 1) = 10^{-2}$. We perform $J := 100$ repeated runs for each of the considered settings for $N := 5 \cdot 10^4$ time steps and thus, obtain a family of trajectories $(\theta^{(j,n)})_{n=1,\dots,N, j=1,\dots,J}$. In each case, we choose the initial values $V(0) := 0$ and the $\theta^{(j,0)} := (0.5, \dots, 0.5)$.

We study the distance of the estimated polynomial to the true function Θ by the relative error:

$$\mathrm{rel\_err}_{n,j} := \frac{\sum_{l=1}^{L} \left(\Theta(x_l) - \sum_{k=1}^{K} \theta_k^{(j,n)} \ell_k(x_l)\right)^2}{\sum_{l'=1}^{L} \Theta(x_{l'})^2},$$

for trajectory $j \in \{1, \dots, J\}$ and time step $n \in \{1, \dots, N\}$. Here $(x_l)_{l=1}^{L}$ are $L := 10^3$ equispaced points in $[-1, 1]$. Moreover, we compare the estimated polynomial to the true function Θ by

$$\mathrm{abs\_err}_{j,x} := \left|\Theta(x) - \sum_{k=1}^{K} \theta_k^{(j,N)} \ell_k(x)\right|$$

for trajectory $j \in \{1, \dots, J\}$ at position $x \in [-1, 1]$. In each case, we study mean and standard deviation (StD) computed over the 100 runs.

Results and discussion

For the polynomial regression problem we now study:

- stochastic gradient descent, as given in (2), with constant learning rate $\eta_{(\cdot)} = h = 0.1$ (Figure 4 top row),
- stochastic gradient descent algorithm, for which the forward Euler update is replaced by an implicit midpoint rule update, with constant learning rate $\eta_{(\cdot)} = h = 0.1$ (Figure 4 bottom row),

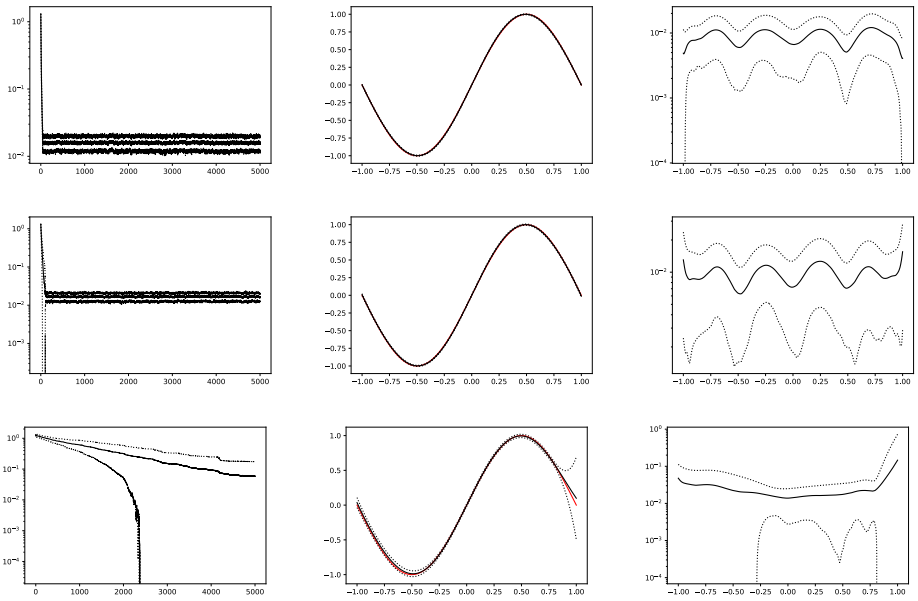

Figure 5: Estimation results of the polynomial regression problem using the stochastic gradient process with reflected Brownian motion process with $\sigma = 5$ (top row), $\sigma = 0.5$ (centre row), and $\sigma = 0.05$ (bottom row). The figures depict the mean over 100 runs (black solid line), mean $\pm$ standard deviation (black dotted line). Left column: trajectory of the rel_err over time; centre column: comparison of Θ (solid red line) and estimated polynomial; right column: estimation error in terms of abs_err.

- the stochastic gradient process with reflected Brownian motion as an index process with standard deviation $\sigma \in \{5, 0.5, 0.05\}$ (Figure 5),
- the stochastic gradient process with Markov pure jump process as an index process with rate parameter $\lambda \in \{10, 1, 0.1, 0.01\}$ (Figure 6), and
- the quality of discretisation of the index processes $(V_t)_{t\geq 0}$ and their convergence (Figure 7).

In addition to those plots, we give means and standard deviations of the relative errors at the terminal state of the iterations in Table 1. To compare the convergence behavior of the different methods, we plot the rel_err within the first 2000 discrete time steps in Figure 8.

We learn several things from these results. Unsurprisingly, the index processes with a strong autocorrelation ($\lambda = 0.01, \sigma = 0.05$) lead to larger errors in the reconstruction: the processes move too slowly to capture the index spaces appropriately, see also Figure 7. In the other cases, we can assume that the processes have reached their stationary regime. Thus, in the figures and table, we should learn about the implicit regularization that is implicated by the different subsampling schemes, see Ali et al. (2020); Smith et al. (2021). We especially see that the mean errors are reduced as σ respectively λ increases, which

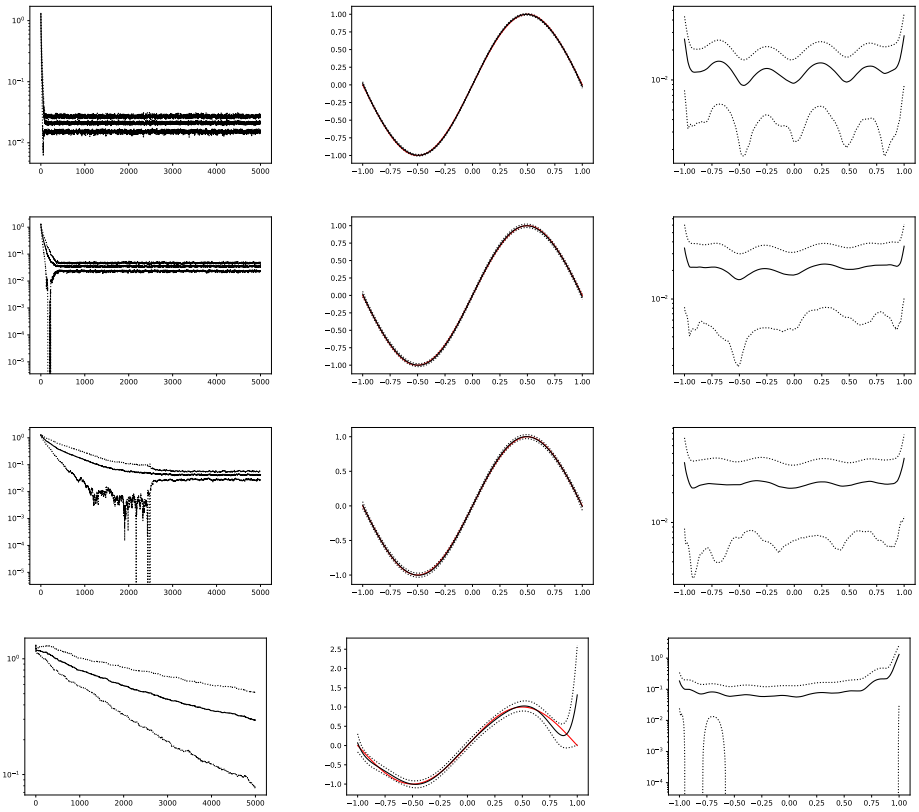

Figure 6: Estimation results of the polynomial regression problem using the stochastic gradient process with pure jump index process with $\lambda = 10$ (first row), $\lambda = 1$ (second row), $\lambda = 0.1$ (third row), and $\lambda = 0.01$ (fourth row). The figures depict the mean over 100 runs (black solid line), mean $\pm$ standard deviation (black dotted line). Left column: trajectory of the rel_err over time; centre column: comparison of Θ (solid red line) and estimated polynomial; right column: estimation error in terms of abs_err.

illustrates the approximation of the full gradient flow as shown in Theorem 3. Although, we should note that we compute the error to the truth Θ, which is likely not the true minimizer of the full optimization problem 31.

It appears that the stochastic gradient processes with reflected diffusion index process $\sigma \in \{0.5, 5\}$ returns the best results. Looking at the error plots in the right column of Figure 5, we see that SGPC especially outperforms the other algorithms close to the boundary. For $\sigma = 5$ this could be seen as a numerical artefact due to the time step $t - t(\cdot - 1)$ being too large. This though is likely not the case for $\sigma = 0.5$, where we see a similar effect, albeit a bit weaker. Indeed, considering Figure 7, we see that the discretised processes oversample the boundary considerably.

In the convergence plot, Figure 8, we see for different methods different speeds of convergence to their respective stationary regime. Those speeds again depend on the autocor-

Method	**Parameters**	**Mean of** $\mathrm{rel\_err}_{N,(\cdot)}$	$\pm$ **StD**
SGD	$\eta_{(\cdot)} = 0.1$	$1.849 \cdot 10^{-2}$	$\pm 5.114 \cdot 10^{-3}$
SGD implicit	$\eta_{(\cdot)} = 0.1$	$1.772 \cdot 10^{-2}$	$\pm 4.483 \cdot 10^{-3}$
SGPC with reflected diffusion index process	$\sigma = 5$	$1.502 \cdot 10^{-2}$	$\pm 3.339 \cdot 10^{-3}$
	$\sigma = 0.5$	$1.627 \cdot 10^{-2}$	$\pm 3.729 \cdot 10^{-3}$
	$\sigma = 0.05$	$5.933 \cdot 10^{-2}$	$\pm 1.153 \cdot 10^{-1}$
SGPC with Markov pure jump index process	$\lambda = 10$	$2.127 \cdot 10^{-2}$	$\pm 5.494 \cdot 10^{-3}$
	$\lambda = 1$	$3.542 \cdot 10^{-2}$	$\pm 1.102 \cdot 10^{-2}$
	$\lambda = 0.1$	$4.193 \cdot 10^{-2}$	$\pm 1.351 \cdot 10^{-2}$
	$\lambda = 0.01$	$2.959 \cdot 10^{-1}$	$\pm 2.183 \cdot 10^{-1}$

Table 1: Accuracy of the estimation in the polynomial regression model. Mean and standard deviation of the relative error of the methods at the final point of their trajectory. In particular, sample mean and sample standard deviation of $j \mapsto \mathrm{rel\_err}_{N,j}$, with $N = 5 \cdot 10^4$, computed over 100 independent runs.

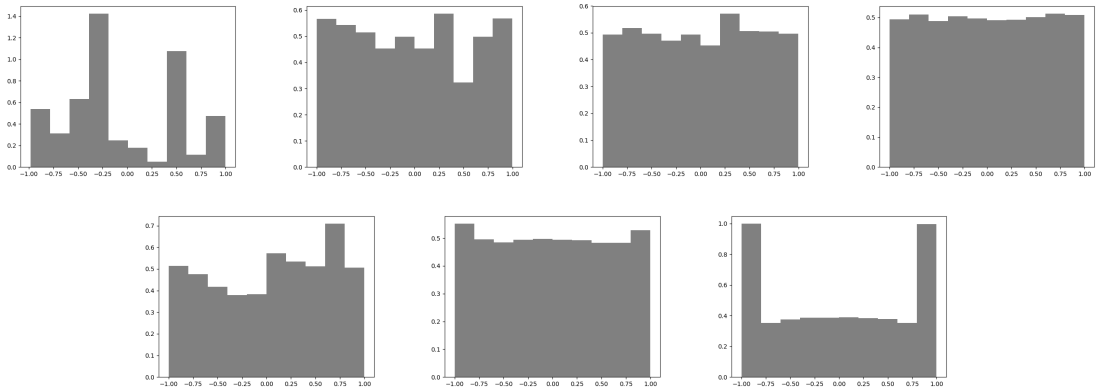

Figure 7: Histograms of single simulations of the index process $(V_t)_{t \geq 0}$, given by the pure jump process (top row) with $\lambda \in \{0.01, 0.1, 1, 10\}$ and the reflected diffusion process (bottom row) with $\sigma \in \{0.05, 0.5, 5\}$, increasing from left to right, respectively. In each case, we simulated $5 \cdot 10^4$ time steps.

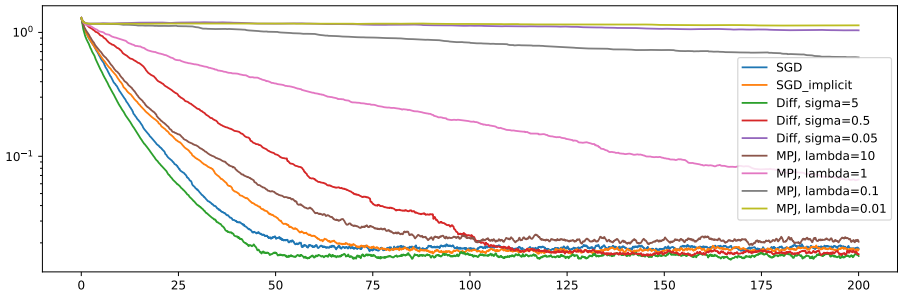

Figure 8: Comparison of the mean rel_err of the stochastic methods for time $t \leq 200$.

relation of the processes. Interestingly, the SGPC with reflected diffusion index process and $\sigma = 5$ appears to be the best of the algorithms.

6.2 Solving partial differential equations using neural networks (NNs)

Partial differential equations (PDEs) are used in science and engineering to model systems and processes, such as: turbulent flow, biological growth, or elasticity. Due to the implicit nature of a PDE and its complexity, the model they represent usually needs to be approximated ('solved') numerically. Finite differences, elements, and volumes have been the state of the art for solving PDEs for the last decades. Recently, deep learning approaches have gained popularity for the approximation of PDE solutions. Here, deep learning is particularly successful in high-dimensional settings, where classical methods suffer from the curse of dimensionality. See for example Raissi et al. (2019); Lu et al. (2021) for physics-informed neural networks (PINNs). Integrated PyTorch-based packages are available for example see Chen et al. (2020); Pedro et al. (2019). More recently, see Li et al. (2021) for state-of-the-art performance results based on the Fourier neural operator.

Physics-informed neural networks are a very natural field of application of deep learning with continuous data. Below we introduce PINNs, the associated continuous-data optimization problem, and the state-of-the-art in the training of PINNs. Then we consider a particular PDE, showcase the applicability of SGP, and compare its performance with the standard SGD-type algorithm.

The basic idea of PINNs consists in representing the PDE solution by a deep neural network where the parameters of the network are chosen such that the PDE is optimally satisfied. Thus, the problem is reduced to an optimization problem with the loss function formulated from differential equations, boundary conditions, and initial conditions. More precisely, for PDE problems of Dirichlet type, we aim to solve a system of equations of type

$$\begin{cases} \mathcal{L}(u(t,x)) &= s(t,x) & (t \in [0,\infty),\ x \in D) \\ u(0,x) &= u_0(x) & (x \in D) \\ u(t,x) &= b(t,x) & (t \in [0,\infty),\ x \in D) \end{cases} \tag{33}$$

where $D \subset \mathbb{R}^d$ is an open, connected, and bounded set and $\mathcal{L}$ is a differential operator defined on a function space V (e.g. $H^1(D)$). The unknown is $u : \bar{D} \to \mathbb{R}^n$. Functions

$s(t,x)$, $b(t,x)$, and $u_0(x)$ are given. In numerical practice, we need to replace the infinite-dimensional space V by a – in some sense – discrete representation. Traditionally, one employs a finite-dimensional subspace of V, say $\mathrm{span}\{\psi_1,\dots,\psi_K\}$, where $\psi_1,\dots,\psi_K$ are basis functions in a finite element method. To take advantage of the recent development of machine learning, one could solve the problem on a set of deep neural networks contained in V, say

$$\Big\{\psi(\cdot;\theta) : \psi(x;\theta) = (W^{(K)}\sigma(\cdot) + b^{(K)}) \circ \cdots \circ (W^{(1)}\sigma(x) + b^{(1)}), x \in [0,\infty) \times D,$$
$$\theta = \Big((W^{(K)}, b^{(K)}), \dots, (W^{(1)}, b^{(1)})\Big) \in \prod_{k=1}^{K} \left(\mathbb{R}^{n_k \times n_{k-1}} \times \mathbb{R}^{n_k}\right) =: X\Big\},$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, applied component-wise, $n_0 = d+1$ and $n_K = 1$ to match input and output of the PDE's solution space, and $n_1,\dots,n_{K-1}$ determine the network's architecture.

In simpler terms, let $u(\cdot;\theta) \in V$ be the output of a feedforward neural network (FNN) with parameters (biases/weights) denoted by $\theta \in X$. The parameters can be learned by minimizing the mean squared error (MSE) loss

$$\Phi(\theta;\mathcal{L},s,u_0,b) := \int_0^\infty w(t) \int_D \left(\mathcal{L}(u(t,x;\theta)) - s(x)\right)^2 \mathrm{d}x\mathrm{d}t + \int_{\partial D} \left(u(0,x;\theta) - u_0(x)\right)^2 \mathrm{d}x$$
$$+ \int_0^\infty w(t) \int_{\partial D} \left(u(t,x;\theta) - b(t,x)\right)^2 \mathrm{d}x\mathrm{d}t,$$

where the first term is the L^2 norm of the PDE residual, the second term is the L^2 norm of the residual for the initial condition, the third term is the L^2 norm of the residual for the boundary conditions, and $w : [0,\infty) \to [0,\infty)$ is an appropriate weight function. The FNN then represents the solution via solving the following minimization problem

$$\min_{\theta \in X} \Phi(\theta;\mathcal{L},s,u_0,b). \tag{34}$$

Note that in physics-informed neural networks, differential operators w.r.t. the input x and the gradient w.r.t the parameter θ are both obtained using automatic differentiation.

Training of physics-informed neural networks

In practice, the optimization problem (34) is often replaced by an optimization problem with discrete potential

$$\widehat{\Phi}(\theta;\mathcal{L},s,u_0,b) := \sum_{k=1}^{K} \left(\mathcal{L}(u(t_k,x_k;\theta)) - s(x_k)\right)^2 + \sum_{k'=1}^{K'} \left(u(0,x'_{k'};\theta) - u_0(x'_{k'})\right)^2$$
$$+ \sum_{k''=1}^{K''} \left(u(t''_{k''},x''_{k''};\theta) - b(t''_{k''},x''_{k''})\right)^2,$$

for appropriate continuous indices

$$(x_k,t_k)_{k=1}^{K} \in [0,\infty)^K \times D^K, (x'_{k'})_{k'=1}^{K'} \in \partial D^K, (x''_{k''},t''_{k''})_{k''=1}^{K''} \in [0,\infty)^K \times \partial D^K$$

that may be chosen deterministically or randomly, see for example Pedro et al. (2019); Lu et al. (2021).

Focusing the training on a fixed set of samples can be problematic: fixing a set of random samples might be unreliable; a reliable cover of the domain will likely only be reached through tight meshing, which scales badly. Sirignano and Spiliopoulos (2018) propose to use SGD on the continuous data space. They employ the discrete dynamic in (2). Naturally, we would like to follow Sirignano and Spiliopoulos (2018) and employ the SGP dynamic on the continuous index set.

To train the PINNs with SGP, we again choose the reflected Brownian motion as an index process, which we discretize with the Euler–Maruyama scheme in Algorithm 2. In addition, we employ mini-batching to reduce the variance in the estimator: We sample $M \in \mathbb{N}$ independent index processes $(V_t^{(1)})_{t\geq 0}, \dots, (V_t^{(M)})_{t\geq 0}$ and then employ the dynamical system

$$\mathrm{d}\theta_t = -\frac{1}{M}\sum_{m=1}^{M} \nabla_\theta f(\theta_t, V_t^{(m)})\mathrm{d}t.$$

Hence, rather than optimizing with respect to a single data set, we optimize with respect to M different data sets in each iteration. While we only briefly mention the mini-batching throughout our analysis, one can easily see that it is fully contained in our framework.

In preliminary experiments, we noticed that the Brownian motion for the sampling on the boundary is not very effective: possibly due to its localizing effect. Hence, we obtain training data on the boundary by sampling uniformly, which we consider justified as a mesh on the boundary scales more slowly as a mesh in the interior and as the boundary behavior of the considered PDE is rather predictable.

PDE and results

We now describe the partial differential equation that we aim to solve with our PINN model. After introducing the PDE we immediately outline the PINN's architectures and show our estimation results. We the train networks on Google Colab Pro using GPUs (often T4 and P100, sometimes K80). We are certain that a more efficient PDE solution could be obtained by classical methods, e.g., the finite element method. We do not compare the deep learning methods with classical methods, as we are mainly interested in SGP and SGD in non-convex continuous-data settings. Other methods that could approximate the PDE solution are not our focus.

The PDE we study is a transport equation; which is a linear first order, time-dependent model. One of the main advantages of studying this particular model is that we know an analytical solution that allows us to compute a precise test error.

Example 1 (1D Transport equation) *We solve the one-dimensional transport equation on the space* $[0, 1]$ *with periodic boundary conditions:*

$$\begin{cases} u_t + u_x &= 0 & (t \in [0,\infty), \quad x \in [0,1]) \\ u(0,x) &= \sin(2\pi x) & (x \in [0,1]) \\ u(t,0) &= u(t,1) & (t \in [0,\infty)). \end{cases} \tag{35}$$

The neural network approximation of this PDE has already been studied by Pedro et al. (2019), our experiments partially use the code associated to this work. The network architecture is defined by a three-layer deep neural network with 128 neurons per layer and a Rectified Linear Unit (ReLu) activation function. While theoretically the solution exists globally in time, we restrict t to a compact domain and w.l.o.g, we assume $t \in [0, 1]$. From the interior of the domain of time and space variables, i.e. $(0, 1) \times (0, 1)$, we use Algorithm 2 with $\sigma = 0.5$ to sample the train set of size $3 \cdot 10^4$ for SGPC and SGPD and we uniformly sample 600 points for the train set of SGD. In addition, as a part of the train set for all three methods, we sample uniformly 20 and 60 points for the initial condition and periodic boundary condition, respectively.

The learning rate for SGD and SGPC is 0.01. The learning rate for SGPD is defined as

$$\eta(t) = \frac{0.01}{\log(t+2)^{0.3}},$$

which is chosen such that the associated $\mu := 1/\eta$ satisfies Assumption 4. For all three methods, we use Adam (see Kingma and Ba, 2015) as the optimizer to speed up the convergence; we use an L^2 regularizer with weight 0.1 to avoid overfitting. Each model is trained over 600 iterations with batch size 50. The training process for SGPC and SGPD contains only one epoch, while we train 50 epochs in the SGD case. We evaluate the models by testing on a uniformly sampled test set of size $2 \cdot 10^3$ and compare the predicted values with the theoretical solution

$$u(t, x) = \sin(2\pi(x - t)).$$

We obtain the losses, the predicted solutions, and the test errors by averaging over 30 random experiments, i.e. 30 independent runs of SGD, SGPC, and SGPD, respectively. We give the results in Figures 9, 10, and 11. Note that the timings are very similar for each of the algorithms, the fact that SGPC and SGPD require us to first sample reflected Brownian motions is negligible.

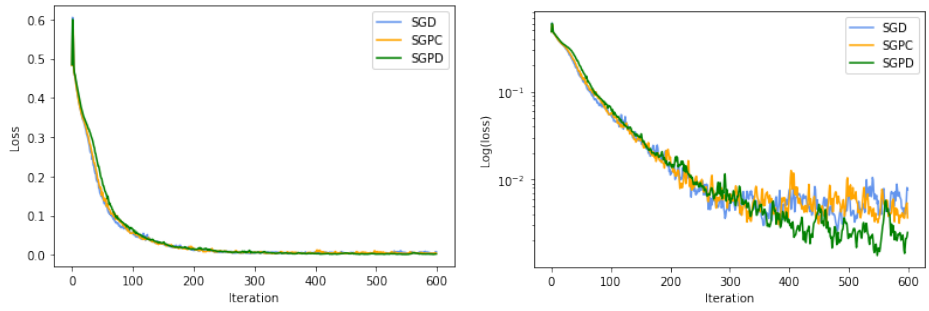

Figure 9: The plots of the loss vs iteration and its log scale for SGD, SGPC, and SGPD. The losses are obtained by averaging over 30 random experiments.

From Figure 9, we notice that while SGD and SGPC behave similarly, SGPD does converge faster. Here, Assumption 4 provides a way of designing a non-constant learning rate in practice. On the test set, the mean squared errors for SGD, SGPC, and SGPD are

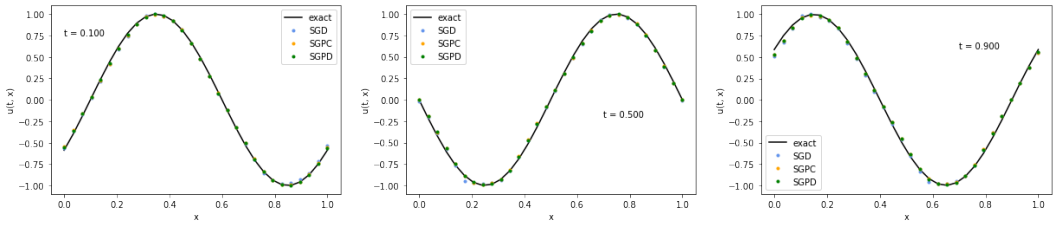

Figure 10: The plots of the solutions at time $t = 0.1, 0.5, 0.9$. We evaluate the models at 30 uniformly sampled points. For each method, the predicted values are taken by averaging over the predicted values from the best models (the model that achieves the lowest training loss within the 600 iteration steps) in 30 random experiments. The black curve is the theoretical solution.

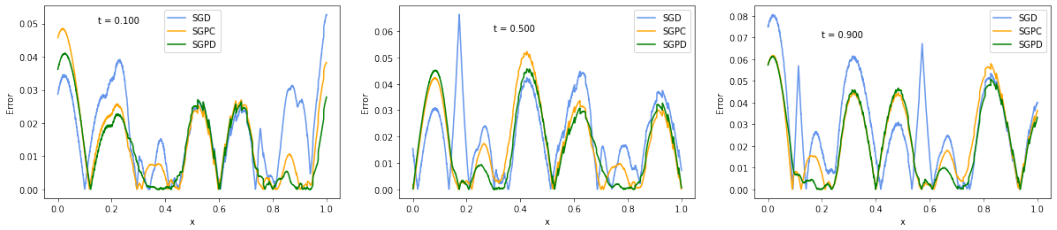

Figure 11: The plots of the test error at time $t = 0.1, 0.5, 0.9$. We evaluate the models at 2000 uniformly sampled points. For each method, the predicted values are taken by averaging over the predicted values from the best models (the model that achieves the lowest training loss within the 600 iteration steps) in 30 random experiments. At each point x, the error calculated by taking the absolute value of the difference between the predicted value and the true solution.

$4.5 \cdot 10^{-4}$, $3.5 \cdot 10^{-4}$, and $2.8 \cdot 10^{-4}$. These test errors refer to the averaged model output of the 30 models from independent experiments. Combined with Figure 10 and Figure 11, we observe that SGPC and SGPD generalize at least slightly better on the test set. This improved generalization error might be due to the additional test data generated by the Brownian motion, as compared to the fixed training set used in PINNs. The combination with the reduction of the learning rate in SGPD, appears to be especially effective.

7. Conclusions and outlook

In this work we have proposed and analyzed a continuous-time stochastic gradient descent method for optimization with respect to continuous data. Our framework is very flexible: it allows for a whole range of random sampling patterns on the continuous data space, which is particularly useful when the data is streamed or simulated. Our analysis shows ergodicity of the dynamical system under convexity assumptions – converging to a stationary measure when the learning rate is constant and to the minimizer when the learning rate decreases.

In experiments we see the suitability of the method and the effect of different sampling patterns on its implicit regularization.

We end this work by now briefly listing some interesting problems for future research in this area. First, we would like to learn how the SGP sampling patterns perform in large-scale (adversarially-)robust machine learning and in other applications we have mentioned but not studied here. Moreover, from a both practical and analytical perspective, it would be interesting to also consider non-compact index spaces S. Those appear especially in robust optimal control and variational Bayes. Finally, we consider the following generalization of the optimization problem (1) to be of high interest:

$$\min_{\theta \in X} \int_S f(\theta, y) \Pi(\mathrm{d}y|\theta),$$

where Π is now a Markov kernel from X to S. Hence, in this case the probability distribution and the sampling pattern itself depend on the parameter θ. Optimization problems of this form appear in the optimal control of random systems (e.g., Deqing and Sergei, 2015) and empirical Bayes (e.g., Casella, 2001) but also in reinforcement learning (e.g., Sutton and Barto, 2018).

Acknowledgments

CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, EPSRC grants EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. JL and CBS acknowledge support from the EPSRC grant EP/S026045/1. Most of the work on this manuscript was done whilst JL was affiliated with the University of Cambridge and Heriot-Watt University.

References

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 233–244, 2020.

Lars Nørvang Andersen, Søren Asmussen, Peter W. Glynn, and Mats Pihlsgård. Lévy processes with two-sided reflection. In *Lévy Matters V: Functionals of Lévy Processes*, pages 67–182. Springer International Publishing, 2015.

Pascal Bianchi. A stochastic proximal point algorithm: convergence and application to convex optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–4, 2015.

Jose Blanchet and Karthyek Murthy. Exact simulation of multidimensional reflected brownian motion. *Journal of Applied Probability*, 55(1):137–156, 2018.

Kristian Bredies and Dirk Lorenz. *Variational Methods*, pages 251–443. Springer International Publishing, 2018.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8:231–357, 2015.

Lucien Le Cam. Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171, 1990.

George Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 12 2001.

Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

Feiyu Chen, David Sondak, Pavlos Protopapas, Marios Mattheakis, Shuheng Liu, Devansh Agarwal, and Marco Di Giovanni. Neurodiffeq: A python package for solving differential equations with neural networks. *Journal of Open Source Software*, 5(46), 2020.

Badr-Eddine Cherief-Abdellatif. Consistency of elbo maximization for model selection. In *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, volume 96, pages 11–31. PMLR, 2019.

Bertrand Cloez and Martin Hairer. Exponential ergodicity for Markov processes with random switching. *Bernoulli*, 21(1):505–536, 2015.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 1310–1320, 2019.

Jana de Wiljes, Sebastian Reich, and Wilhelm Stannat. Long-time stability and accuracy of the ensemble kalman–bucy filter for fully observed processes and small measurement noise. *SIAM Journal on Applied Dynamical Systems*, 17(2):1152–1181, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

Huang Deqing and Chernyshenko Sergei. Long-time average cost control of stochastic systems using sum of squares of polynomials. In *2015 34th Chinese Control Conference (CCC)*, pages 2344–2349, 2015.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Armin Eftekhari, Bart Vandereycken, Gilles Vilmart, and Konstantinos C. Zygalakis. Explicit stabilised gradient descent for faster strongly convex optimisation. *Bit Numer Math*, 61:119–139, 2021.

Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Per Christian Hansen. *Discrete Inverse Problems*. Society for Industrial and Applied Mathematics, 2010.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*, 2015.

Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021.

Harold Kushner. *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, volume 6 of *MIT Press Series in Signal Processing, Optimization, and Control*. MIT Press, Cambridge, 1984.

Harold Kushner. *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. Birkhäuser Basel, 1990.

Harold Kushner and George Yin. *Stochastic Approximation Algorithms and Recursive Algorithms and Applications*. Springer, New York, NY, 2003.

Jonas Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31(39), 2021.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 2101–2110, 2017.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations, ICLR*, 2021.

Thomas Liggett. *Continuous Time Markov Processes: An Introduction*. American Mathematical Soc., 2010.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Yingjie Liu. Discretization of a class of reflected diffusion processes. *Mathematics and Computers in Simulation*, 38(1):103–108, 1995.

Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, 2014.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 2016)*, pages 354–363, 2016.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Sandra May, Rolf Rannacher, and Boris Vexler. Error analysis for a finite element approximation of elliptic dirichlet boundary control problems. *SIAM Journal on Control and Optimization*, 51(3):2585–2611, 2013.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Juan B. Pedro, Juan Maronas, and Roberto Paredes. Solving partial differential equations with neural networks. *arXiv:1912.04737*, 2019.

Roger Pettersson. Approximations for stochastic differential equations with reflecting convex boundaries. *Stochastic Processes and their Applications*, 59:295–308, 1995.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, page 2817–2826, 2017.

Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical Solution of Ordinary Differential Equations*, pages 479–538. Springer Berlin Heidelberg, 2007.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Daniel Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Heidelberg, 2013.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer New York, 2004.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019.

Beatriz Sinova, Gil González-Rodríguez, and Stefan Van Aelst. M-estimators of location for functional data. *Bernoulli*, 24(3):2328–2357, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time. *SIAM Journal on Financial Mathematics*, 8(1):933–961, 2017.

Justin Sirignano and Konstantinos Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations, ICLR*, 2021.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Nicolas Garcia Trillos and Daniel Sanz-Alonso. The Bayesian Update: Variational Formulations and Gradient Flows. *Bayesian Analysis*, 15(1):29–56, 2020.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part II: Continuous time analysis. *arXiv:2106.02588*, 2021.