Continuous-time stochastic gradient descent for optimizing over the stationary distribution of stochastic differential equations

Ziheng Wang  |  Justin Sirignano

Abstract
We develop a new continuous-time stochastic gradient descent method for optimizing over the stationary distribution of stochastic differential equation (SDE) models. The algorithm continuously updates the SDE model’s parameters using an estimate for the gradient of the stationary distribution. The gradient estimate is simultaneously updated using forward propagation of the SDE state derivatives, asymptotically converging to the direction of steepest descent. We rigorously prove convergence of the online forward propagation algorithm for linear SDE models (i.e., the multidimensional Ornstein–Uhlenbeck process) and present its numerical results for nonlinear examples. The proof requires analysis of the fluctuations of the parameter evolution around the direction of steepest descent. Bounds on the fluctuations are challenging to obtain due to the online nature of the algorithm (e.g., the stationary distribution will continuously change as the parameters change). We prove bounds for the solutions of a new class of Poisson partial differential equations (PDEs), which are then used to analyze the parameter fluctuations in the algorithm. Our algorithm is applicable to a range of mathematical finance applications involving statistical calibration of SDE models and stochastic optimal control for long time horizons where ergodicity of
the data and stochastic process is a suitable modeling framework. Numerical examples explore these potential applications, including learning a neural network control for high-dimensional optimal control of SDEs and training stochastic point process models of limit order book events.

**KEYWORDS**
machine learning, online optimization, Poisson equations, stochastic differential equations, stochastic gradient descent

## 1 | INTRODUCTION

Consider a parametric process \( X_t^\theta \in \mathbb{R}^d \) which satisfies the stochastic differential equation (SDE):

\[
\begin{align*}
    dX_t^\theta &= \mu(X_t^\theta, \theta) \, dt + \sigma(X_t^\theta, \theta) \, dW_t, \\
    X_0^\theta &= x,
\end{align*}
\]

where \( \theta \in \mathbb{R}^\ell, \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}^{d \times d}, \) and \( W_t \) is a standard Brownian motion. Suppose \( X_t^\theta \) is ergodic with the stationary distribution \( \pi^\theta. \)

Our goal is to select the parameters \( \theta \) which minimize the objective function

\[
J(\theta) = \sum_{n=1}^{N} \left( \mathbb{E}_{\pi^\theta}[f_n(Y)] - \beta_n \right)^2,
\]

where \( Y \) is a random variable with distribution \( \pi^\theta, f_n \) are known functions, and \( \beta_n \) are the target quantities. Thus, we are interested in optimizing the parameterized SDEs (1) such that their stationary distribution matches, as closely as possible, the target statistics \( \beta_n. \) In practice, the target statistics may be data from real-world observations which are then used to calibrate the SDE model (1).

### 1.1 | Existing methods to optimize over the stationary distribution of SDEs

The stationary distribution \( \pi^\theta \) is typically unknown and therefore it is challenging to optimize over \( J(\theta). \) The quantity \( \mathbb{E}_{Y \sim \pi^\theta}[f_n(Y)] \) as well as its gradient \( \nabla_{\theta} \mathbb{E}_{Y \sim \pi^\theta}[f_n(Y)] \) must be estimated in

---

1 Sufficient conditions (Pardoux, Veretennikov, 2003) for the existence and uniqueness of \( \pi^\theta \) are: (1) both coefficients \( \mu \) and \( \sigma \) are assumed to be bounded and \( \sigma \) is uniformly continuous with respect to \( x \) variable, (2) \( \lim_{|x| \to \infty} \sup_{\theta} |\mu(x, \theta)| = -\infty, \) and (3) there exist two constants \( 0 < \lambda < \Lambda < \infty \) such that \( \lambda I_d \leq \sigma \sigma^T(x, \theta) \leq \Lambda I_d \) where \( I_d \) is the \( d \times d \) identity matrix.
order to minimize $J(\theta)$. $E_{Y \sim \pi_\theta}[f_n(Y)]$ can be evaluated using the forward Kolmogorov equation

$$\mathcal{L}^\theta_x p_\infty(x, \theta) = 0,$$

(3)

where $\mathcal{L}^\theta_x$ is the infinitesimal generator of the process $X^\theta_t$ and $\mathcal{L}^\theta_x$, $\mathcal{L}_x^\theta E_{Y \sim \pi_\theta}[f_n(Y)]$ can be calculated using an appropriate adjoint partial differential equation (PDE) for Equation (3) (Annunziato & Borzì, 2013; Butt, 2022; Fleig & Guglielmi, 2017; Kaltenbacher & Pedretscher, 2018). However, if the dimension of $d$ for $X^\theta_t$ is large, solving the forward Kolmogorov equation and its adjoint PDE become extremely computationally expensive. In the special case where the drift function $\mu$ is the gradient of a scalar function and the volatility function $\sigma$ is constant, there exists a closed-form formula for the stationary distribution (Pavliotis, 2014).

Alternatively, $E_{Y \sim \pi_\theta}[f_n(Y)]$ can be approximated by simulating Equation (1) over a long time $[0, T]$. Similar to Carmona and Laurière (2023), the gradient descent algorithm would be

- Simulate $X^\theta_t$ for $t \in [0, T]$.
- Evaluate the gradient of $J_T(\theta) := \sum_{n=1}^N \left( \frac{1}{T} \int_0^T f_n \left( X^\theta_t \right) dt - \beta_n \right)^2$.
- Update the parameter as $\theta_{k+1} = \theta_k - \alpha_k \nabla_\theta J_T(\theta_k)$,

where $\alpha_k$ is the learning rate. This gradient descent algorithm will be slow; a long simulation time $T$ will be required for each optimization iteration. A second disadvantage is that $J_T(\theta)$ is an approximation to $J(\theta)$ and therefore error is introduced into the algorithm, that is, $\nabla_\theta J_T(\theta) \neq \nabla_\theta J(\theta)$.

### 1.2 An online optimization algorithm

We propose a new continuous-time stochastic gradient descent (SGD) algorithm, which allows for computationally efficient optimization of Equation (2). The algorithm uses **online forward propagation** to asymptotically estimate the gradient of the objective function with respect to the parameters. For notational convenience (and without loss of generality), we will set $N = 1$ and $\beta_1 = \beta$. The online forward propagation algorithm for optimizing Equation (2) is

$$\frac{d\theta_t}{dt} = -2\alpha_t (f(\tilde{X}_t) - \beta)(\nabla f(X_t)\tilde{X}_t)^\top,$$

$$dX_t = (\nabla_x \mu(X_t, \theta_t)X_t + \nabla_\theta \mu(X_t, \theta_t))dt + (\nabla_x \sigma(X_t, \theta_t)X_t + \nabla_\theta \sigma(X_t, \theta_t))dW_t,$$

$$d\tilde{X}_t = \mu(\tilde{X}_t, \theta_t)dt + \sigma(\tilde{X}_t, \theta_t)dW_t,$$

where $W_t$ and $\tilde{W}_t$ are independent Brownian motions and $\alpha_t$ is the learning rate. Before proceeding with our analysis, we, first, clarify the notation in Equation (4). In this paper, the Jacobian matrix of a vector value function $f : x \in \mathbb{R}^n \rightarrow \mathbb{R}^m$ is an $m \times n$ matrix, that is, $\nabla_x f(x) \in \mathbb{R}^{n \times m}$. When the function has only one variable, we may omit the subscript in the gradient. For example, we may use $\nabla f(x)$ to denote $\nabla_x f(x)$. For functions of several variables, we use the subscript...
in the gradient to denote the partial derivative with respect to a subset of variables. For example, we will use \( \nabla_x \mu(X_t^\theta, \theta) \) to denote \( \nabla_x \mu(x, \theta)|_{x=X_t^\theta} \). Therefore, the variables have the following dimensions:

\[
\hat{X}_t \in \mathbb{R}^{d \times \ell}, \quad \nabla_x \mu \in \mathbb{R}^{d \times d}, \quad \nabla_\theta \mu \in \mathbb{R}^{d \times \ell}, \quad \nabla_x \sigma \in \mathbb{R}^{d \times d \times d}, \quad \nabla_\theta \sigma \in \mathbb{R}^{d \times d \times \ell}.
\]

Let \( \hat{X}_t^i \) denote the \( i \)-th row of \( \hat{X}_t \), and then the dynamics of \( \hat{X}_t \) in Equation (4) are

\[
d\hat{X}_t^i = (\nabla_x \mu_i(X_t, \theta)) \hat{X}_t^i + \nabla_\theta \mu_i(X_t, \theta))dt + \sum_{j=1}^d (\nabla_x \sigma_{i,j}(X_t, \theta)) \hat{X}_t^i + \nabla_\theta \sigma_{i,j}(X_t, \theta))dW_t^j.
\]

In Equation (4), \( \hat{X}_t \) and \( X_t \) have the same dynamics, although they are driven by independent Brownian motions. The role of \( \hat{X}_t \) will be explained in detail later in this section. The learning rate \( \alpha_t \) in Equation (4) must be chosen such that \( \int_0^\infty \alpha_s ds = \infty \) and \( \int_0^\infty \alpha_s^2 ds < \infty \). (An example is \( \alpha_t = \frac{C}{1+t^d} \)). \( \hat{X}_t \) estimates the derivative of \( X_t \) with respect to \( \theta \). The parameter \( \theta_i \) is continuously updated using \((f(\hat{X}_t) - \beta)(\nabla f(X_t)X_t)^\top \) as a stochastic estimate for \( \nabla_\theta J(\theta) \). Deterministic gradient descent in continuous-time is often referred to as a “gradient flow”; therefore, the proposed algorithm can be viewed as a “stochastic gradient flow.”

To better understand the algorithm (4), let us first rewrite the gradient of the objective function using the ergodicity of \( X_t^\theta \):

\[
\nabla_\theta J(\theta) = 2 \left( E_{Y \sim \pi_0} f(Y) - \beta \right) \nabla_\theta E_{Y \sim \pi_0} f(Y) \quad \text{a.s.}
\]

\[
= 2 \left( \lim_{T \to \infty} \frac{1}{T} \int_0^T f\left(X_t^\theta\right) dt - \beta \right) \cdot \nabla_\theta \left( \lim_{T \to \infty} \frac{1}{T} \int_0^T f\left(X_t^\theta\right) dt \right). \quad (5)
\]

If the derivative and the limit can be interchanged, the gradient can be expressed as

\[
\nabla_\theta J(\theta) = 2 \left( \lim_{T \to \infty} \frac{1}{T} \int_0^T f\left(X_t^\theta\right) dt - \beta \right) \cdot \lim_{T \to \infty} \frac{1}{T} \int_0^T \nabla f\left(X_t^\theta\right) \nabla_\theta X_t^\theta dt. \quad (6)
\]

Define \( \hat{X}_t^\theta = \nabla_\theta X_t^\theta \) and, under mild regularity conditions for the coefficients (see for example Röckner et al., 2021; Wang & Sirignano, 2022), \( \hat{X}_t^\theta \) will satisfy

\[
d\hat{X}_t^\theta = (\nabla_x \mu\left(X_t^\theta, \theta\right) \hat{X}_t^\theta + \nabla_\theta \mu\left(X_t^\theta, \theta\right))dt + (\nabla_x \sigma\left(X_t^\theta, \theta\right) \hat{X}_t^\theta + \nabla_\theta \sigma\left(X_t^\theta, \theta\right))dW_t. \quad (7)
\]

Note that \( \hat{X}_t \) and \( \hat{X}_t^\theta \) satisfy the same equations, except \( \theta \) is a fixed constant for \( \hat{X}_t^\theta \) while \( \theta_i \) is updated continuously in time for \( \hat{X}_t \). Then, we have that

\[
\nabla_\theta J(\theta) = 2 \left( \lim_{T \to \infty} \frac{1}{T} \int_0^T f\left(X_t^\theta\right) dt - \beta \right) \cdot \lim_{T \to \infty} \frac{1}{T} \int_0^T \nabla f\left(X_t^\theta\right) \hat{X}_t^\theta dt. \quad (8)
\]

The formula (8) can be used to evaluate \( \nabla_\theta J(\theta) \) and thus allows for optimization via a gradient descent algorithm. However, as highlighted in Section 1.1, \( X_t^\theta \) must be simulated for a large time period \([0, T]\) for each optimization iteration, which is computationally costly. A natural
alternative is to develop a continuous-time SGD algorithm which updates $\theta$ using a stochastic estimate $G(\theta_t)$ for $\nabla_\theta J(\theta_t)$, where $G(\theta_t)$ asymptotically converges to an unbiased estimate for the direction of steepest descent $\nabla_\theta J(\theta_t)$. (The random variable $G(\theta_t)$ is called an unbiased estimate for $\nabla_\theta J(\theta_t)$ if $\mathbb{E}[G(\theta_t) | \theta_t] = \nabla_\theta J(\theta_t)$.) The online algorithm (4) does exactly this using $G(\theta_t) = 2(\bar{f}(\bar{X}_t) - \beta)\nabla f(X_t)\bar{X}_t$ as a stochastic estimate for $\nabla_\theta J(\theta_t)$.

For large $t$, we expect that $\mathbb{E}[\bar{f}(\bar{X}_t) - \beta] \approx \mathbb{E}_{Y \sim \pi_{\theta}}[f(Y) - \beta]$ and $\mathbb{E}[\nabla f(X_t)\bar{X}_t] \approx \nabla_\theta \mathbb{E}_{Y \sim \pi_{\theta}}[f(Y) - \beta]$ since $\theta_t$ is changing very slowly as $t$ becomes large due to $\lim_{t \to \infty} \alpha_t = 0$. Here we highlight that for random variables $X$ and $Y$, it is not typically true that $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$ unless $X$ and $Y$ are independent. This is the reason why the process $\bar{X}_t$ is introduced. Since $\bar{X}_t$ and $X_t$ are driven by independent Brownian motions, we expect that $\mathbb{E}[2(\bar{f}(\bar{X}_t) - \beta)\nabla f(X_t)\bar{X}_t] \approx \nabla_\theta J(\theta_t)$ for large $t$ due to $\bar{X}_t$ and $(X_t, \bar{X}_t)$ becoming asymptotically independent since $\theta_t$ will be changing very slowly for large $t$. Thus, we expect that for large $t$, the stochastic sample $G(\theta_t) = 2(\bar{f}(\bar{X}_t) - \beta)\nabla f(X_t)\bar{X}_t$ will provide an asymptotically unbiased estimate for the direction of steepest descent $\nabla_\theta J(\theta_t)$ and $\|\nabla_\theta J(\theta_t)\|$ will converge to zero as $t \to \infty$.

### 1.3 Contributions of this paper

We rigorously prove the convergence of the algorithm (4) when $\mu(\cdot)$ is linear and for constant $\sigma$. Even in the linear case, the distribution of $(X_t, \bar{X}_t, \bar{X}_t, \theta_t)$ will be non-Gaussian and convergence analysis is non-trivial. Unlike in the traditional SGD algorithm, the data are not i.i.d. (i.e., $X_t$ is correlated with $X_s$ for $s \neq t$) and, for a finite time $t$, the stochastic update direction $G(\theta_t)$ is not an unbiased estimate of $\nabla_\theta J(\theta_t)$. One must show that asymptotically $G(\theta_t)$ becomes an unbiased estimate of the direction of steepest descent $\nabla_\theta J(\theta_t)$. Furthermore, it must be proven that the stochastic fluctuations of $G(\theta_t)$ around the direction of steepest descent vanish in an appropriate way as $t \to \infty$.

The proof, therefore, requires analysis of the fluctuations of the stochastic update direction $G(\theta_t)$ around $\nabla_\theta J(\theta_t)$. Bounds on the fluctuations are challenging to obtain due to the online nature of the algorithm. The stationary distribution $\pi_{\theta_t}$ will continuously change as the parameters $\theta_t$ evolve. We prove bounds on a new class of Poisson PDEs, which are then used to analyze the parameter fluctuations in the algorithm. The fluctuations are rewritten in terms of the solution to the Poisson PDE using Ito’s lemma, the PDE solution bounds are subsequently applied, and then we can show asymptotically that the fluctuations vanish. Our main theorem proves for the multidimensional Ornstein–Uhlenbeck process that

$$\lim_{t \to \infty} \|\nabla_\theta J(\theta_t)\| \overset{a.s.}{=} 0. \tag{9}$$

In the numerical section of this paper, we evaluate the performance of our online algorithm (4) for a variety of linear and nonlinear examples. In these examples, we show that the algorithm can also perform well in practice for nonlinear SDEs. We also demonstrate that the online algorithm can optimize over path-dependent SDEs and pathwise statistics of SDEs such as the auto-covariance. In addition, we also demonstrate the applications of the online optimization algorithm to mathematical finance problems, such as SDE model calibration, parameter estimation for partially observed SDEs, high-dimensional stochastic control problems, and limit order book models.
1.4 Literature review

In this paper, we show that, if \( \alpha_t \) is appropriately chosen, then \( \nabla \theta J(\theta_t) \to 0 \) as \( t \to \infty \) with probability 1. Similar results have been previously proven for SGD in discrete time. Bertsekas and Tsitsiklis (2000) prove the convergence of SGD with i.i.d. data samples. Benveniste et al. (2012) prove the convergence of SGD in discrete time with the correlated data samples under stronger conditions than Bertsekas and Tsitsiklis (2000). We refer readers to Benveniste et al. (2012), Bertsekas and Tsitsiklis (2000), Bottou et al. (2018), Goodfellow et al. (2016), Kushner and Yin (2003) for a thorough review of the very large literature on SGD and similar stochastic optimization algorithms (e.g., SGD with momentum, Adagrad, ADAM, and RMSprop). However, these articles do not study SGD methods for optimizing over the stationary distribution of stochastic models, which is the focus of our paper. Forward propagation algorithms have been used for optimizing recurrent neural networks, where the algorithm is referred to as “real time recurrent learning” (Benzing et al., 2019; Menick et al., 2018; Williams & Zipser, 1989).

Recent articles such as Bhudisaksang and Cartea (2021), Sharrock and Kantas (2023), Sirignano and Spiliopoulos (2017, 2020), Surace and Pfister (2018) have studied continuous-time SGD. Sirignano and Spiliopoulos (2017) proposed a “stochastic gradient descent in continuous time” (SGDCT) algorithm for estimating parameters \( \theta \) in an SDE \( X_t^\theta \) from continuous observations of \( X_t^{\theta_*} \) where \( \theta_* \) is the true parameter. Sirignano and Spiliopoulos (2017) prove convergence of the algorithm to a stationary point. Bhudisaksang and Cartea (2021) extended SGDCT to estimate the drift parameter of a continuous-time jump-diffusion process. Sirignano and Spiliopoulos (2020) analyzed proved a central limit theorem for the SGDCT algorithm and a convergence rate for strongly convex objective functions. Sharrock and Kantas (2023) established the almost sure convergence of two-timescale SGD algorithms in continuous time. Surace and Pfister (2018) designed an online learning algorithm for estimating the parameters of a partially observed diffusion process and studied its convergence. Sharrock et al. (2021) propose an online estimator for the parameters of the McKean–Vlasov SDE and proves that this estimator converges in \( L_1 \) to the stationary points of the asymptotic log-likelihood.

Our paper has several important differences as compared to Bhudisaksang and Cartea (2021), Sharrock and Kantas (2023), Sharrock et al. (2021), Sirignano and Spiliopoulos (2017, 2020), Surace and Pfister (2018). These previous papers estimate the parameter \( \theta \) for the SDE \( X_t^\theta \) from observations of \( X_t^{\theta_*} \) where \( \theta_* \) is the true parameter. In this paper, our goal is to select \( \theta \) such that the stationary distribution of \( X_t^\theta \) matches certain target statistics. Therefore, unlike the previous papers, we are directly optimizing over the stationary distribution of \( X_t^\theta \). The presence of the \( X \) process in SGDCT makes the mathematical analysis challenging as the \( X \) term introduces correlation across times, and this correlation does not disappear as time tends to infinity. In order to prove convergence, Sirignano and Spiliopoulos (2017, 2020) use an appropriate Poisson PDE (Gilbarg & Trudinger, 2015; Pardoux & Veretennikov, 2001, 2003) associated with \( X \) to describe the evolution of the parameters for large times and analyze the fluctuations of the parameter around the direction of steepest descent. However, the theoretical results from Pardoux and Veretennikov (2001, 2003) do not apply to the PDE considered in this paper since the diffusion term in our PDE is not uniformly elliptic. This is a direct result of the process \( X_t \) in Equation (4), which shares the same Brownian motion with the process \( X_t \). In the case of constant \( \sigma \), the PDE operator will not be uniformly elliptic and, furthermore, the coefficient for derivatives such as \( \frac{\partial^2}{\partial x^2} \) is zero. Consequently, we must analyze a new class of Poisson PDEs, which is different than the class of Poisson PDEs studied in Pardoux and Veretennikov (2001, 2003). We prove that there exists a solution to
this new class of Poisson PDEs, which satisfies polynomial bounds. The polynomial bounds are crucial for analyzing the fluctuations of the parameter evolution in the algorithm (4).

1.5 Organization of paper

The paper is organized into three main sections. In Section 2, we present the assumptions and the main theorem. Section 3 rigorously proves the convergence of our algorithm for multidimensional linear SDEs. Section 4 studies the numerical performance of our algorithm for a variety of linear and nonlinear SDEs, including McKean–Vlasov and path-dependent SDEs. Applications of the online optimization algorithm in mathematical finance are discussed, including SDE model calibration, parameter estimation for partially observed SDE models, stochastic optimal control, and mean-field games. Numerical examples demonstrate how the method can be used to numerically solve high-dimensional stochastic optimal control problems and high-dimensional stochastic models of limit order book events.

2 MAIN RESULT

In this section, we rigorously prove convergence of the algorithm (4) for the following multidimensional Ornstein–Uhlenbeck process:

$$\begin{align*}
  dX^\theta_t &= (g(\theta) - h(\theta)X^\theta_t)dt + \sigma dW_t,
  
  X^\theta_0 &= x,
\end{align*}$$

where \( \theta \in \mathbb{R}^d \), \( g(\theta) \in \mathbb{R}^d \), \( h(\theta) \in \mathbb{R}^{d \times d}_+ \), \( W_t \in \mathbb{R}^d \), \( X^\theta_t \in \mathbb{R}^d \), and \( \sigma \) is a scalar constant. Since \( h(\theta) \) is positive definite, the solution to the SDE (10) is

$$X^\theta_t = e^{-h(\theta)t}x + (h(\theta))^{-1}(Id - e^{-h(\theta)t})g(\theta) + e^{-h(\theta)t}\int_{0}^{t}e^{-h(\theta)s}\sigma dW_s,$$

where \( Id \) is the \( d \times d \) identity matrix. Let \( \pi_\theta \) be the stationary distribution of \( X^\theta_t \). \( \pi_\theta \) exists and is unique; for example, see Pavliotis, 2014.) Our goal is to solve the optimization problem

$$\min_{\theta} J(\theta) = \min_{\theta} \left( \mathbb{E}_{Y \sim \pi_\theta} f(Y) - \beta \right)^2,$$

where \( \beta \) is a constant. To solve Equation (12), our online algorithm (4) becomes

$$\begin{align*}
  \frac{d\theta_t}{dt} &= -2\alpha_t(f(\tilde{X}_t) - \beta)\nabla f(X_t)\tilde{X}_t, \\
  dX_t &= (g(\theta_t) - h(\theta_t)X_t)dt + \sigma dW_t, \\
  \frac{dX_t}{dt} &= \nabla_\theta g(\theta_t) - \nabla_\theta h(\theta_t)X_t - h(\theta_t)\tilde{X}_t, \\
  d\tilde{X}_t &= (g(\theta_t) - h(\theta_t)\tilde{X}_t)dt + \sigma dW_t,
\end{align*}$$

(13)
where $W_t$ and $\hat{W}_t$ are independent Brownian motions, $\nabla_{\theta} g(\theta_t) \in \mathbb{R}^{d \times \ell}$, $\nabla_{\theta} h(\theta_t) \in \mathbb{R}^{d \times d \times \ell}$ and $X_t \in \mathbb{R}^{d \times \ell}$ is the gradient process for $X_t$. The element $(i, j)$ of the process $\hat{X}_t$ satisfies:

$$
\frac{d}{dt} \hat{X}_t^{i,j} = \frac{\partial g_i(\theta)}{\partial \theta_j} - \sum_{k=1}^{d} \frac{\partial h_{ik}(\theta)}{\partial \theta_j} X_t^k - \sum_{k=1}^{d} h_{ik}(\theta) \hat{X}_t^{k,j}, \quad i \in \{1, 2, \ldots, d\}, \quad j \in \{1, 2, \ldots, \ell\}. \quad (14)
$$

For the rest of this article, we will use $C, C_k, C_p$ to denote generic constants. Our convergence theorem will require the following assumptions.

**Assumption 2.1.**

1. $g(\theta), \nabla_{\theta}^i g(\theta), h(\theta)$, and $\nabla_{\theta}^j h(\theta)$ are uniformly bounded functions for $i = 1, 2$.
2. $h$ is symmetric and uniformly positive definite, that is, there exists a constant $c > 0$ such that

$$
\min \{ x^T h(\theta) x \} \geq c|x|^2, \quad \forall \theta \in \mathbb{R}^\ell, x \in \mathbb{R}^d.
$$

3. $f, \nabla^i f, i = 1, 2, 3$ are polynomially bounded:

$$
|f(x)| + \sum_{i=1}^{3} |\nabla^i f(x)| \leq C(1 + |x|^m^\hat{\theta}), \quad \forall x \in \mathbb{R}^d
$$

for some constant $C, m^\hat{\theta} > 0$.

4. The learning rate $\alpha_t$ satisfies $\int_0^\infty \alpha_t \, dt = \infty$, $\int_0^\infty \alpha_t^2 \, dt < \infty$, $\int_0^\infty |\alpha'_t| \, ds < \infty$, and there is a $\hat{p} > 0$ such that $\lim_{t \to \infty} \alpha_t^{1/2 + 2\hat{p}} = 0$.

Under these assumptions, we are able to prove the following convergence result.

**Theorem 2.2.** Under Assumption 2.1 and for the Ornstein–Uhlenbeck process (10), the algorithm (13) will converge to a stationary point almost surely:

$$
\lim_{t \to \infty} |\nabla_{\theta} J(\theta_t)| = 0. \quad (16)
$$

### 3 PROOF OF THEOREM 2.2

In this section, we present the proof of Theorem 2.2. We begin by decomposing the evolution of $\theta_t$ in Equation (13) into several terms:

$$
\frac{d\theta_t}{dt} = -2\alpha_t(f(\hat{X}_t) - \beta)(\nabla f(X_t) \hat{X}_t)^T
$$

$$
= -2\alpha_t(E_{Y \sim \pi_{\theta_t}} f(Y) - \beta)(\nabla f(X_t) \hat{X}_t)^T - 2\alpha_t \left( f(\hat{X}_t) - E_{Y \sim \pi_{\theta_t}} f(Y) \right) (\nabla f(X_t) \hat{X}_t)^T
$$

$|\cdot|$ denotes the Euclidean norm. Sometimes for a square matrix $x$, $|x|$ will be used to denote its spectral norm, which is equivalent to the Euclidean norm.
\[
\frac{d\theta_t}{dt} = -\alpha_t \nabla \log f(\theta_t) - 2\alpha_t (E_{Y \sim \pi_{\theta_t}} f(Y) - \beta) (\nabla f(X_t) X_t - \nabla \theta E_{Y \sim \pi_{\theta_t}} f(Y))^T
\]

Direction of steepest descent

Fluctuation term 1

\[-2\alpha_t \left( f(X_t) - E_{Y \sim \pi_{\theta_t}} f(Y) \right) (\nabla f(X_t) X_t)^T \]

Fluctuation term 2

(17)

Define the error terms

\[Z_1^i = (E_{Y \sim \pi_{\theta_t}} f(Y) - \beta) (\nabla f(X_t) X_t - \nabla \theta E_{Y \sim \pi_{\theta_t}} f(Y))^T, \]

\[Z_2^i = \left( f(X_t) - E_{Y \sim \pi_{\theta_t}} f(Y) \right) (\nabla f(X_t) X_t)^T. \]

(18)

We have, therefore, decomposed the evolution of \( \theta_t \) into the direction of steepest descent \(-\alpha_t \nabla \log f(\theta_t)\) and the two fluctuation terms \(2\alpha_t Z_1^i\) and \(2\alpha_t Z_2^i\).

As in Sirignano and Spiliopoulos (2017), we study a cycle of stopping times to control the time periods where \(|\nabla \log f(\theta_t)|\) is close to zero and away from zero. Let us select an arbitrary constant \( \kappa > 0 \) and also define \( \mu = \mu(\kappa) > 0 \) (to be chosen later). Then set \( \sigma_0 = 0 \) and define the cycles of random times

\[0 = \sigma_0 \leq \tau_1 \leq \sigma_1 \leq \tau_2 \leq \sigma_2 \leq \ldots, \]

where for \( k = 1, 2, \ldots \)

\[\tau_k = \inf \{ t > \sigma_{k-1} : |\nabla \log f(\theta_t)| \geq \kappa \} \]

\[\sigma_k = \sup \left\{ \frac{|\nabla \log f(\theta_s)|}{2} : \nabla \log f(\theta_s) \leq 2 |\nabla \log f(\theta_{\tau_k})| \text{ for all } s \in [\tau_k, t] \text{ and } \int_{\tau_k}^t \alpha_s ds \leq \mu \right\}. \]

(19)

We define the random time intervals \( J_k = [\sigma_{k-1}, \tau_k) \) and \( I_k = [\tau_k, \sigma_k) \). We introduce \( \eta > 0 \), which will be chosen to be sufficiently small later. We first seek to control

\[\Delta_{t_k, \sigma_k + \eta}^i := \int_{t_k}^{\sigma_k + \eta} \alpha_s Z_s^i ds, \quad i = 1, 2 \]

(20)

and, as in Sirignano and Spiliopoulos (2017), we will use a Poisson equation to bound the online fluctuation terms \(\Delta_{t_k, \sigma_k + \eta}^i\) where the ergodic properties of \( X^\theta_t \) will be leveraged in the analysis.

In this paper, we focus on the Ornstein–Uhlenbeck process (10). As in Equation (7), its gradient process \( X^\theta_t := \nabla \theta X^\theta_t = \left( \frac{\partial^2}{\partial \theta_j} \right)_{i, j} \in \mathbb{R}^{d \times \ell} \) now satisfies the SDE:

\[\frac{dX^\theta_t}{dt} = \nabla g(\theta) - \nabla h(\theta) X^\theta_t - h(\theta) X^\theta_t, \quad (21)\]
which can be equivalently written as

\[
\frac{d}{dt} \frac{\partial X_t^{\theta}}{\partial \theta_j} = \frac{\partial g_i(\theta)}{\partial \theta_j} - \sum_{k=1}^{d} \frac{\partial h_{ik}(\theta)}{\partial \theta_j} X_t^{\theta,k} - \sum_{k=1}^{d} h_{ik}(\theta) \frac{\partial X_t^{\theta,k}}{\partial \theta_j},
\]

for \( i \in \{1, 2, \ldots, d\} \) and \( j \in \{1, 2, \ldots, \ell\} \). Thus, we know the solution of Equation (21) with initial point \( \tilde{x} \) is

\[
X_t^{\theta} = e^{-h^{(\theta)}(\tilde{x})} \tilde{x} + e^{-h^{(\theta)}(x)} \int_0^t e^{h^{(\theta)}(s)} \left( \nabla_{\theta} g(\theta) - \nabla_{\theta} h^{(\theta)}(x) X_s^{\theta} \right) ds.
\]

The independent Ornstein-Uhlenbeck process used to obtain the asymptotic unbiased gradient is

\[
\frac{d}{dt} \hat{X}_t^{\theta} = (g(\theta) - h(\theta) \hat{X}_t^{\theta})dt + \sigma d\hat{W}_t,
\]

where \( \hat{W}_t \) is another Brownian motion independent of \( W_t \). For the processes \( X_t^{\theta}, \tilde{X}_t^{\theta}, \hat{X}_t^{\theta} \) in Equations (10), (21), and (24), we can prove the following convergence results.

**Proposition 3.1.** Let \( p_t(x, x', \theta) \) and \( p_\infty(x', \theta) \) denote the transition probability and invariant density of the multidimensional Ornstein–Uhlenbeck process (10). Under Assumption 2.1, we have the following ergodic result:

(i) For any \( m > 0 \), there exists a constant \( C = C(m) \) such that

\[
\left| \nabla^{i}_\theta p_\infty(x', \theta) \right| \leq \frac{C}{1 + |x'|^m}, \quad i = 0, 1, 2.
\]

(ii) For any \( m', k \), there exist constants \( C, m \) such that for any \( t > 1 \)

\[
\left| \nabla^{i}_\theta p_t(x, x', \theta) - \nabla^{i}_\theta p_\infty(x', \theta) \right| \leq \frac{C(1 + |x|^m)}{(1 + |x'|^{m'})^k(1 + t)^k}, \quad i = 0, 1, 2.
\]

(iii) For any \( m', k \), there exist constants \( C, m \) such that for any \( t > 1 \)

\[
\left| \nabla^{i}_x \nabla^{j}_\theta p_t(x, x', \theta) \right| \leq \frac{C(1 + |x|^m)}{(1 + |x'|^{m'})^k(1 + t)^k}, \quad i = 0, 1, \quad j = 1, 2.
\]

(iv) For any \( m > 0 \), there exists a constant \( C = C(m) \) such that for any \( t \geq 0 \)

\[
E_x \left| X_t^{\theta} \right|^m \leq C(1 + |x|^m), \quad E_{x, \tilde{x}} \left| \tilde{X}_t^{\theta} \right|^m \leq C(1 + |x|^m + |\tilde{x}|^m).
\]
Here $E_x$ denotes that the initial condition for the process $X^\theta_t$ is $x$, that is, $X^\theta_0 = 0$. $E_{x,\bar{x}}$ denotes that the initial conditions of the processes $(X^\theta_t, \bar{X}^\theta_t)$ in Equation (7) are $(x, \bar{x})$, that is, $X^\theta_0 = x$ and $\bar{X}^\theta_0 = \bar{x}$.

(v) For any function $f$ satisfying Equation (15), there exists constants $C, m$ such that for any $t \in [0,1]$

\[
\left| \nabla^i_x \nabla^j_\theta E_x f(X^\theta_t) \right| \leq C(1 + |x|^m), \quad i = 0, 1, \quad j = 0, 1, 2.
\]  

(29)

Remark 3.2. Proposition 3.1 is similar to Theorem 1 in Pardoux and Veretennikov (2003). However, the assumption of uniform boundedness in Pardoux and Veretennikov (2001) does not hold for the multidimensional Ornstein–Uhlenbeck process (10). Thus, we give a brief proof by direct calculations in Section A.

We must analyze the fluctuation terms $Z^1_t$ and $Z^2_t$. In order to do this, we prove a polynomially bounded solution exists to a new class of Poisson PDEs. The polynomial bound is in the spatial coordinates and, importantly, the bound is uniform in the parameter $\theta$. A Poisson PDE was also used in Sirignano and Spiliopoulos (2017). However, several key innovations are required for the online optimization algorithm (13) that we consider in this paper. Unlike in Sirignano and Spiliopoulos (2017), $X^\theta_t$ in Equation (21) does not have a diffusion term, which means that $(X^\theta_t, \bar{X}^\theta_t)$ is a degenerate diffusion process and its generator $\mathcal{L}^\theta_{x,\bar{x}}$ is not a uniformly elliptic operator. Thus, we cannot use the results from Pardoux and Veretennikov (2001, 2003). Instead, we must prove existence and bounds for this new class of Poisson PDEs.

Lemma 3.3. Define the error function

\[
G^1(x, \bar{x}, \theta) = (E_{Y \sim \pi_\theta} f (Y) - \beta) \left( \nabla f (x) \bar{x} - \nabla_\theta E_{Y \sim \pi_\theta} f (Y) \right)^T
\]  

and

\[
v^1(x, \bar{x}, \theta) = - \int_0^\infty E_{x,\bar{x}} G^1(X^\theta_t, \bar{X}^\theta_t, \theta) \, dt,
\]  

where $E_{x,\bar{x}}$ is a conditional expectation given $X^\theta_0 = x$ and $\bar{X}^\theta_0 = \bar{x}$. Then, under Assumption 2.1, $v^1(x, \bar{x}, \theta)$ is the classical solution of the Poisson equation

\[
\mathcal{L}^\theta_{x,\bar{x}} u(x, \bar{x}, \theta) = G^1(x, \bar{x}, \theta),
\]  

where $u = (u_1, \ldots, u_\ell)^T \in \mathbb{R}^\ell$ is a vector, $\mathcal{L}^\theta_{x,\bar{x}} u(x, \bar{x}, \theta) = (\mathcal{L}^\theta_{x,\bar{x}} u_1(x, \bar{x}, \theta), \ldots, \mathcal{L}^\theta_{x,\bar{x}} u_\ell(x, \bar{x}, \theta))^T$, and $\mathcal{L}^\theta_{x,\bar{x}}$ is the infinitesimal generator of the process $(X^\theta_t, \bar{X}^\theta_t)$, that is, for any test function $\varphi$

\[
\mathcal{L}^\theta_{x,\bar{x}} \varphi(x, \bar{x}) = \mathcal{L}^\theta_{x} \varphi(x, \bar{x}) + \text{tr} \left( \nabla_\bar{x} \varphi(x, \bar{x}) (\nabla_\theta g(\theta) - \nabla_\theta h(\theta) x - h(\theta) \bar{x}) \right).
\]  

(33)
Furthermore, there exist an integer \( m' \) and a constant \( C = C(m') \), which do not depend upon \((x, \tilde{x}, \vartheta)\) such that the solution \( v^1 \) satisfies the bound

\[
|v^1(x, \tilde{x}, \vartheta)| + |\nabla_{\vartheta} v^1(x, \tilde{x}, \vartheta)| + |\nabla_{x} v^1(x, \tilde{x}, \vartheta)| + |\nabla_{\tilde{x}} v^1(x, \tilde{x}, \vartheta)| \leq C \left( 1 + |x|^{m'} + |\tilde{x}|^{m'} \right). \tag{34}
\]

The proof of Lemma 3.3 is in Appendix B. We will next study the fluctuation terms \( Z_i \). It will be necessary to prove bounds on the moments of \( X_t \) and \( \tilde{X}_t \) in order to analyze the error term \( \Delta_{i, \omega_k + \eta}^i \).

**Lemma 3.4.** For any \( p > 0 \), there exists a constant \( C_p \) that only depends on \( p \) such that the processes \( X_t, \tilde{X}_t \) from Equation (13) satisfy

\[
E_x|X_t|^p \leq C_p(1 + |x|^p), \quad E_{x, \tilde{x}}|\tilde{X}_t|^p \leq C_p(1 + |x|^p + |\tilde{x}|^p). \tag{35}
\]

Furthermore, we have the bounds

\[
E_x \left( \sup_{0 \leq t' \leq t} |X_{t'}|^p \right) = O(\sqrt{t}) \quad \text{as } t \to \infty, \tag{36}
\]

\[
E_{x, \tilde{x}} \left( \sup_{0 \leq t' \leq t} |\tilde{X}_{t'}|^p \right) = O(\sqrt{t}) \quad \text{as } t \to \infty.
\]

**Proof.** By adapting the method in Giles and Fang (2020), we first prove Equation (35) for \( p \geq 2 \) and then the result for \( 0 < p < 2 \) follows from Hölder’s inequality. Let \( p = 2m \) and applying Itô’s formula to \( e^{p \alpha t} |X|^{2m} \), we have for any \( t \geq 0 \),

\[
e^{p \alpha t/2} |X_t|^p - |X_0|^p \leq \int_0^t p \left( \frac{\alpha}{2} |X_s|^2 + \langle X_s, g(\vartheta_s) - h(\vartheta_s)X_s \rangle \right) e^{p \alpha s/2} |X_s|^{p-2} \, ds
\]

\[
+ \int_0^t \frac{p(p-1)d}{2} e^{p \alpha s/2} |X_s|^{p-2} + \int_0^t pe^{p \alpha s/2} |X_s|^{p-2} \langle X_s, dW_s \rangle,
\]

where \( \langle a, b \rangle := a^T b \). By Assumption 2.1, we know that there exists constants \( \alpha > 0, \beta > d \) such that for any \( \vartheta \)

\[
\langle x, g(\vartheta) - h(\vartheta)x \rangle \leq -\alpha |x|^2 + \beta. \tag{38}
\]

Thus by taking expectations on both sides of Equation (37) and using Equation (38), we obtain

\[
E_x \left[ e^{p \alpha t/2} |X_t|^p \right] - |x|^p \leq \int_0^t -\frac{p\alpha}{2} E_x \left[ e^{p \alpha s/2} |X_s|^p \right] \, ds + \int_0^t E_x \left[ \frac{p(p+1)\beta}{2} e^{p \alpha s/2} |X_s|^{p-2} \right] \, ds.
\]

Young’s inequality implies that

\[
\frac{p(p+1)\beta}{2} e^{p \alpha s/2} |X_s|^{p-2} \leq \frac{p\alpha}{2} e^{p \alpha s/2} |X_s|^p + c_p e^{p \alpha s/2}
\]
where \( c_p = (\frac{p-2}{p\alpha})^{p/2-1}(\beta(p + 1))^{p/2} \). Therefore, we obtain

\[
\mathbb{E}_x [e^{p\alpha t/2} | X_t |^p] - |x|^p \leq \int_0^t c_pe^{p\alpha s/2} ds
\]

and

\[
\mathbb{E}_x |X_t|^p \leq \frac{2c_p}{p\alpha} + e^{-p\alpha t/2} |x|^p \leq C_p(1 + |x|^p).
\]

Using the moment bound for \( X_t \), we can derive the moment bound for \( \tilde{X}_t \). From Equations (23) and (13), we know

\[
\tilde{X}_t = e^{-\int_0^t h(\theta_u) du} X_0 + \int_0^t e^{-\int_s^t h(\theta_u) du} (\nabla \theta g(\theta_s) - \nabla \theta h(\theta_s) X_s) ds
\]

and thus

\[
\mathbb{E}_{\tilde{X}, \tilde{x}} |\tilde{X}_t|^p \leq 2|\tilde{x}|^p + 2\mathbb{E}_{\tilde{X}, \tilde{x}} \int_0^t \left| e^{-\int_s^t h(\theta_u) du} \right| |\nabla \theta g(\theta_s) - \nabla \theta h(\theta_s) X_s| ds \leq (a) \left| \tilde{x}\right|^p + C_p \mathbb{E}_{\tilde{X}, \tilde{x}} \int_0^t \left| e^{-\int_s^t h(\theta_u) du} \right| (1 + |\tilde{x}|) ds
\]

\[
\leq 2|\tilde{x}|^p + C_p \left( \mathbb{E}_{\tilde{X}} \int_0^t \left| e^{-\int_s^t h(\theta_u) du} \right| (1 + |\tilde{x}|) ds \right)^p \leq 2|\tilde{x}|^p + C_p \left( \mathbb{E}_{\tilde{X}} \int_0^t \left| e^{-\int_s^t h(\theta_u) du} \right| \right)^p (b) \leq 2|\tilde{x}|^p + C_p \left( \mathbb{E}_{\tilde{X}} \int_0^t \left| e^{-\int_s^t h(\theta_u) du} \right| \right)^p \leq C_p (1 + |x|^p + |\tilde{x}|^p)
\]

where step (a) is by Assumption 2.1 and the fact

\[
\lambda_{\text{max}} \left( e^{-\int_s^t h(\theta_u) du} \right) = e^{-\lambda_{\text{min}} \int_s^t h(\theta_u) du} \leq e^{-c(t-s)}
\]

and step (b) is by Jensen’s inequality.

To prove Equation (36), we use a similar method as in Pardoux and Veretennikov (2001). By Itô’s formula, we have for \( p \geq 1 \)

\[
|X_t|^p - |X_0|^p \leq \int_0^t 2p|X_s|^{2p-2} \langle X_s, g(\theta_s) - h(\theta_s) X_s \rangle ds
\]

\[
+ \int_0^t (p + 2p - 1)|X_s|^{2p-2} ds + 2p \int_0^t |X_s|^{2p-2} \langle X_s, dW_s \rangle
\]

\[
\leq C_p \int_0^t |X_s|^{2p-2} ds + 2p \int_0^t |X_s|^{2p-2} \langle X_s, dW_s \rangle.
\]
Using the Burkholder–Davis–Gundy inequality, there exists a constant $C$ such that

$$
\mathbb{E}_x \left( \sup_{t' \leq t} |X_{t'}|^{2p} \right) \leq |x|^{2p} + C_p \left( \mathbb{E}_x \int_0^t |X_s|^{p-2} ds \right)^{1/2} + C_p \mathbb{E}_x \int_0^t |X_s|^{2p-2} ds,
$$

(43)

which together with estimate (35) can be used to derive the bound

$$
\mathbb{E}_x \left( \sup_{t' \leq t} |X_{t'}|^2 \right) \leq |x|^2 + C_p \left( \mathbb{E}_x \int_0^t |X_s|^2 ds \right)^{1/2} + C_p \mathbb{E}_x \int_0^t |X_s|^2 ds,
$$

(44)

Furthermore, for $t \geq 1$,

$$
\mathbb{E}_x \left( \sup_{t' \leq t} |X_{t'}|^p \right)^\frac{1}{p} \leq \left( |x|^p + C_p \left( t + t^{1/2} \right) \right)^\frac{1}{p} \leq |x| + C_p \left( 1 + |x|^{p-\frac{1}{2}} \right) \sqrt{t},
$$

(45)

where step (a) is by Hölder inequality. Similarly, we have for any $p' < p$ and $t \geq 1$ that

$$
\mathbb{E}_x \left( \sup_{t' \leq t} |X_{t'}|^{p'} \right) \leq C |x|^{p'} + C \left( 1 + |x|^{p-\frac{1}{2}} \right) t^{p' \over 2p},
$$

(46)

and thus the result for $X_t$ in Equation (36) follows. Finally, similarly as in Equation (40),

$$
\mathbb{E}_{x, \bar{x}} \sup_{t' \leq t} |\bar{X}_{t'}|^p \leq 2 |\bar{x}|^p + 2 \mathbb{E}_{x, \bar{x}} \sup_{t' \leq t} \left| \int_0^{t'} e^{-c(t'-s)} \Delta \eta \left( \partial \theta \right) ds \right| \left| \nabla \theta g(\theta) - \nabla \theta h(\theta) \right| \mathbb{E}_{x, \bar{x}} \left| \sup_{t' \leq t} |X_{t'}|^p \right| ds
$$

$$
\leq 2 |\bar{x}|^p + C_p \mathbb{E}_x \sup_{t' \leq t} \left| \int_0^{t'} e^{-c(t'-s)} (1 + |X_s|) ds \right| \mathbb{E}_{x, \bar{x}} \sup_{t' \leq t} |X_{t'}|^p.
$$

(47)

Combining Equations (45), (46), and (47), we can prove the bound for $\bar{X}_t$ in Equation (36).

Using the estimates in Lemmas 3.3 and 3.4, we can now bound the first fluctuation term $\Delta^1_{\varepsilon_k, \sigma_k + \eta}$ in Equation (20).

**Lemma 3.5.** Under Assumption 2.1, for any fixed $\eta > 0$

$$
\left| \Delta^1_{\varepsilon_k, \sigma_k + \eta} \right| \to 0 \text{ as } k \to \infty, \quad \text{a.s.}
$$

(48)

**Proof.** The idea is to use the Poisson equation in Lemma 3.3 to derive an equivalent expression for the term $\Delta^1_{\varepsilon_k, \sigma_k + \eta}$, which we can appropriately control as $k$ becomes large. Consider the function

$$
\Delta^1(x, \bar{x}, \theta) = \left( \mathbb{E}_{Y \sim \pi_0} f(Y) - \beta \right) \left( \nabla f(x) \bar{x} - \nabla \theta \mathbb{E}_{Y \sim \pi_0} f(Y) \right)^T.
$$
By Lemma 3.3, the Poisson equation $L^\partial_{x,x} u(x, \tilde{x}, \theta) = G^1(x, \tilde{x}, \theta)$ will have a unique smooth solution $v^1(x, \tilde{x}, \theta)$ that grows at most polynomially in $(x, \tilde{x})$. Let us apply Itô’s formula to the function

$$u^1(t, x, \tilde{x}, \theta) := \alpha_i v^1(x, \tilde{x}, \theta) \in \mathbb{R}^\ell,$$

evaluated on the stochastic process $(X_t, \tilde{X}_t, \theta_t)$. Recall that $u_i$ denotes the $i$th element of $u$ for $i \in \{1, 2, \ldots, \ell\}$. Then,

$$u^1_t(\sigma, X_\sigma, \tilde{X}_\sigma, \theta_\sigma) = u^1_t(\tau, X_\tau, \tilde{X}_\tau, \theta_\tau) + \int_\tau^\sigma \tilde{\theta}_s u^1_s(s, X_s, \tilde{X}_s, \theta_s) ds + \int_\tau^\sigma L^\partial_{x,x} u^1_s(s, X_s, \tilde{X}_s, \theta_s) ds$$

$$+ \int_\tau^\sigma \nabla \theta u^1_s(s, X_s, \tilde{X}_s, \theta_s) d\theta_s + \int_\tau^\sigma \nabla \varepsilon u^1_s(s, X_s, \tilde{X}_s, \theta_s) \sigma dW_s.$$ (49)

Rearranging the previous equation, we obtain the representation

$$\Delta^1_{t, \xi, \alpha_\xi + \eta} = \int_{t_k}^{\sigma_{k+1}} \alpha_i G^1(X_s, \tilde{X}_s, \theta_s) ds = \int_{t_k}^{\sigma_{k+1}} L^\partial_{x,x} u^1(s, X_s, \tilde{X}_s, \theta_s) ds$$

$$= \alpha_\xi + \eta v^1(\sigma_{k+1}, \tilde{X}_{\sigma_{k+1}}, \theta_{\sigma_{k+1}}) - \alpha_\xi v^1(\sigma_k, \tilde{X}_{\sigma_k}, \theta_{\sigma_k}) - \int_{t_k}^{\sigma_{k+1}} \alpha_i v^1(X_s, \tilde{X}_s, \theta_s) ds$$

$$+ \int_{t_k}^{\sigma_{k+1}} 2\alpha_i^2 \nabla \varepsilon u^1(X_s, \tilde{X}_s, \theta_s) f(X_s)^T (\nabla f(X_s)) \tau ds - \int_{t_k}^{\sigma_{k+1}} \alpha_i \nabla \varepsilon u^1(X_s, \tilde{X}_s, \theta_s) dW_s.$$ (50)

The next step is to treat each term on the right-hand side of Equation (50) separately. For this purpose, let us first set

$$J^{1,1}_t = \alpha_i \sup_{s \in [0, t]} |v^1(X_s, \tilde{X}_s, \theta_s)|.$$ (51)

By Equations (34) and (36), there exists a constant $C$ that only depends on $m'$ such that

$$\mathbb{E} \left| J^{1,1}_t \right|^2 \leq C \alpha_i^2 \mathbb{E} \left[ 1 + \sup_{s \in [0, t]} |X_s|^{m'} + \sup_{s \in [0, t]} |\tilde{X}_s|^{m'} \right]$$

$$= C \alpha_i^2 \left[ 1 + \sqrt{t} \sup_{s \in [0, t]} |X_s|^{m'} + \sqrt{t} \sup_{s \in [0, t]} |\tilde{X}_s|^{m'} \right]$$

$$\leq C \alpha_i^2 \sqrt{t}.$$ (52)

Let $\rho > 0$ be the constant in Assumption 2.1 such that $\lim_{t \to \infty} \alpha_i^2 t^{1/2 + 2\rho} = 0$ and for any $\delta \in (0, \rho)$ define the event $A_{t, \delta} = \{J^{1,1}_t \geq t^{\delta - \rho}\}$. Then we have for $t$ large enough such that $\alpha_i^2 t^{1/2 + 2\rho} \leq 1$

$$\mathbb{P}(A_{t, \delta}) \leq \frac{\mathbb{E} \left| J^{1,1}_t \right|^2}{t^{2(\delta - \rho)}} \leq C \frac{\alpha_i^2 t^{1/2 + 2\rho}}{t^{2\delta}} \leq C \frac{1}{t^{2\delta}},$$
The latter implies that
\[
\sum_{n \in \mathbb{N}} P(A_{2^n, \delta}) < \infty.
\]

Therefore, by the Borel–Cantelli lemma, we have that for every \( \delta \in (0, p) \) there is a finite positive random variable \( d(\omega) \) and some \( n_0 < \infty \) such that for every \( n \geq n_0 \) one has
\[
J_{2^n}^{1,1} \leq \frac{d(\omega)}{2^{n(p-\delta)}},
\]

Thus, for \( t \in [2^n, 2^{n+1}) \) and \( n \geq n_0 \) one has for some finite constant \( C < \infty \)
\[
J_t^{1,1} \leq C \alpha_{2^{n+1}} \sup_{s \in (0, 2^n]} |v_1^1(X_s, \bar{X}_s, \vartheta_s)| \leq C \frac{d(\omega)}{2^{(n+1)(p-\delta)}} \leq C \frac{d(\omega)}{t^{p-\delta}},
\]

which proves that for \( t \geq 2^{n_0} \) with probability one
\[
J_t^{1,1} \leq C \frac{d(\omega)}{t^{p-\delta}} \to 0, \quad \text{as } t \to \infty. \tag{53}
\]

Next we consider the term
\[
J_{t,0}^{1,2} = \int_0^t \left| \alpha'_s v_1^1(X_s, \bar{X}_s, \vartheta_s) - 2\alpha_s^2 v_0 v_1^1(X_s, \bar{X}_s, \vartheta_s)(f(\bar{X}_s) - \beta)(\nabla f(X_s) \bar{X}_s)^\top \right| ds.
\]

There exists a constant \( 0 < C < \infty \) (that may change from line to line) and \( 0 < m' < \infty \) such that
\[
\sup_{t > 0} \mathbb{E}\left| J_{t,0}^{1,2} \right| \overset{(a)}{\leq} C \int_0^\infty \left( |\alpha'_s| + \alpha_s^2 \right) \left( 1 + \mathbb{E}|X_s|^{m'} + \mathbb{E}|\bar{X}_s|^{m'} + \mathbb{E}|\tilde{X}_t|^{m'} \right) ds
\]
\[
\overset{(b)}{\leq} C \int_0^\infty \left( |\alpha'_s| + \alpha_s^2 \right) ds
\]
\[
\leq C,
\]

where step (a) is by Assumption 2.1 and Equation (34) and in step (b) we use Equation (35). Thus, there is a finite random variable \( J_{\infty,0}^{1,2} \) such that
\[
J_{t,0}^{1,2} \to J_{\infty,0}^{1,2}, \quad \text{as } t \to \infty \text{ with probability one.} \tag{54}
\]

The last term we need to consider is the martingale term
\[
J_{t,0}^{1,3} = \int_0^t \alpha_s \nabla v_1^1(X_s, \bar{X}_s, \vartheta_s) dW_s.
\]
By Doob’s inequality, Assumption 2.1 Equations (34), (35), and using calculations similar to the ones for the term $J_{t,0}^{1,2}$, we can show that for some finite constant $C < \infty$,

$$\sup_{t>0} \mathbb{E}|J_{t,0}^{1,3}|^2 \leq C \int_0^\infty \alpha_t^2 ds < \infty.$$ 

Thus, by Doob’s martingale convergence theorem, there is a square integrable random variable $J_{\infty,0}^{1,3}$ such that

$$J_{t,0}^{1,3} \to J_{\infty,0}^{1,3}, \quad \text{as } t \to \infty \text{ both almost surely and in } L^2.$$ 

(55)

Let us now return to Equation (50). Using the terms $J_{t}^{1,1}, J_{t,0}^{1,2}$, and $J_{t,0}^{1,3}$, we can write

$$|\Delta_{t_k}^{1,1,\sigma_k+\eta}| \leq J_{t_k}^{1,1,\sigma_k+\eta} + J_{t_k}^{1,2,\sigma_k+\eta,\tau_k} + J_{t_k}^{1,3,\sigma_k+\eta,\tau_k},$$

which together with Equations (53)–(55) prove the statement of the lemma. \(\Box\)

Now we prove a similar convergence result for $\Delta_{t_k}^{2,2,\tau_k}$. We first give an extension of Lemma 3.3 for the Poisson equation.

**Lemma 3.6.** Define the error function

$$G^2(x, \bar{x}, \tilde{x}, \theta) = [f(\bar{x}) - \mathbb{E}_{Y \sim \pi_\theta} f(Y)](\nabla f(x) \bar{x})^\top.$$ 

(56)

and

$$v^2(x, \bar{x}, \tilde{x}, \theta) = -\int_0^\infty \mathbb{E}_{X_0, \bar{X}_0, \tilde{X}_0} G^2(X_0^\theta, \bar{X}_0^\theta, \tilde{X}_0^\theta, \theta) \, dt,$$ 

(57)

where $\mathbb{E}_{X_0, \bar{X}_0, \tilde{X}_0}$ is a conditional expectation given $X_0^\theta = x$, $\bar{X}_0^\theta = \bar{x}$, and $\tilde{X}_0^\theta = \tilde{x}$. Under Assumption 2.1, $v^2(x, \bar{x}, \tilde{x}, \theta)$ is the classical solution of the Poisson equation

$$\mathcal{L}^\theta_{x, \bar{x}, \tilde{x}} u(x, \bar{x}, \tilde{x}, \theta) = G^2(x, \bar{x}, \tilde{x}, \theta),$$ 

(58)

where $\mathcal{L}^\theta_{x, \bar{x}, \tilde{x}}$ is generator of the process $(X_0^\theta, \bar{X}_0^\theta, \tilde{X}_0^\theta)$, that is, for any test function $\phi$

$$\mathcal{L}^\theta_{x, \bar{x}, \tilde{x}} \phi(x, \bar{x}, \tilde{x}) = \mathcal{L}^\theta_{x, \bar{x}} \phi(x, \bar{x}, \tilde{x}) + \mathcal{L}^\theta_{\tilde{x}} \phi(x, \bar{x}, \tilde{x}).$$ 

(59)

Furthermore, there exist an integer $m'$ and a constant $C = C(m')$, which do not depend upon $(x, \bar{x}, \tilde{x}, \theta)$ such that the solution $v^2$ satisfies the bound

$$|v^2(x, \bar{x}, \tilde{x}, \theta)| + |\nabla_x v^2(x, \bar{x}, \tilde{x}, \theta)| + |\nabla_{\bar{x}} v^2(x, \bar{x}, \tilde{x}, \theta)| + |\nabla_{\tilde{x}} v^2(x, \bar{x}, \tilde{x}, \theta)|$$

$$\leq C(1 + |x|^{m'} + |\bar{x}|^{m'} + |\tilde{x}|^{m'}).$$ 

(60)

The proof of Lemma 3.6 can be found in Appendix B.
Lemma 3.7. Under Assumption 2.1, for any fixed \( \eta > 0 \), we have

\[
\left| \Delta^2_{\tau_k, \sigma_k + \eta} \right| \to 0, \text{ as } k \to \infty, \quad \text{a.s.} \quad (61)
\]

Proof. Consider the function

\[
G^2(x, \tilde{x}, \theta) = (f(\tilde{x}) - E_{Y \sim \pi_\eta} f(Y))(\nabla f(x) \tilde{x})^\top. \quad (62)
\]

Let \( v^2 \) be the solution of Equation (58) in Lemma 3.6. We apply Itô formula to the function \( u^2(t, x, \tilde{x}, \theta) = \alpha_i v^2(x, \tilde{x}, \theta) \) evaluated on the stochastic process \((X_t, \tilde{X}_t, X_t, \theta_t)\) and get for any

\[
u_i^2(\sigma, \tilde{X}_\sigma, \tilde{X}_\sigma, \theta_\sigma) - u_i^2(\tau, X_T, \tilde{X}_\tau, \theta_\tau) = \int_{\tau}^{\sigma} \delta \sigma u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) ds
\]

\[
+ \int_{\tau}^{\sigma} \mathcal{L}_{X, \tilde{X}, X} u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) ds + \int_{\tau}^{\sigma} \mathcal{L}_{\tilde{X}, X} u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) ds + \int_{\tau}^{\sigma} \nabla_{\tilde{X}} u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) d\theta_s
\]

\[
+ \int_{\tau}^{\sigma} \nabla_{X} u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) dW_s + \int_{\tau}^{\sigma} \nabla_{\tilde{X}} u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) d\tilde{W}_s. \quad (63)
\]

Rearranging the previous equation, we obtain the representation

\[
\Delta^2_{\tau_k, \sigma_k + \eta} = \int_{\tau_k}^{\sigma_k + \eta} \alpha_i G^2(X_s, \tilde{X}_s, X_s, \theta_s) ds = \int_{\tau_k}^{\sigma_k + \eta} \mathcal{L}_{X, \tilde{X}, X} u_i^2(s, X_s, \tilde{X}_s, X_s, \theta_s) ds
\]

\[
= \alpha_{\sigma_k + \eta} u^2(X_{\sigma_k + \eta}, \tilde{X}_{\sigma_k + \eta}, X_{\sigma_k + \eta}, \theta_{\sigma_k + \eta}) - \alpha_i u^2(X_{\sigma_k}, \tilde{X}_{\sigma_k}, X_{\sigma_k}, \theta_{\sigma_k})
\]

\[
- \int_{\tau_k}^{\sigma_k + \eta} \alpha_i u^2(X_s, \tilde{X}_s, X_s, \theta_s) ds - \int_{\tau_k}^{\sigma_k + \eta} \alpha_i \nabla_{\tilde{X}} u^2(X_s, \tilde{X}_s, X_s, \theta_s) dW_s
\]

\[
+ \int_{\tau_k}^{\sigma_k + \eta} 2\alpha_i \nabla_{\tilde{X}} u^2(X_s, \tilde{X}_s, X_s, \theta_s)(f(\tilde{X}_s) - \beta)(\nabla f(X_s) \tilde{x})^\top d\tilde{W}_s - \int_{\tau_k}^{\sigma_k + \eta} \alpha_i \nabla_{\tilde{X}} u^2(X_s, \tilde{X}_s, X_s, \theta_s) dW_s. \quad (64)
\]

The next step is to treat each term on the right-hand side of Equation (64) separately. For this purpose, let us first set

\[
J^{2,1}_t = \alpha_t \sup_{s \in [0, t]} |u^2(X_s, \tilde{X}_s, X_s, \theta_s)|. \quad (65)
\]

Using the same approach as for \( X_t \) in Lemma 3.4, we can show that for any \( p > 0 \), there exists a constant \( C_p \) that only depends on \( p \) such that

\[
E_{\tilde{x}} |X_t|^p \leq C_p (1 + |	ilde{x}|^p), \quad E_{\tilde{x}} \left( \sup_{0 \leq t' \leq t} |X_{t'}|^p \right) = O(\sqrt{t}) \quad \text{as } t \to \infty. \quad (66)
\]
Combining Lemma 3.4 and Equations (60) and (66), we know that there exists a constant $C$ such that

$$
\mathbb{E}[J^{2,1}_t]^2 \leq C \alpha^2 \mathbb{E} \left[ 1 + \sup_{s \in [0,t]} |X_s|^m + \sup_{s \in [0,t]} |\bar{X}_s|^m + \sup_{s \in [0,t]} |\hat{X}_s|^m \right]
$$

$$
= C \alpha^2 \left[ 1 + \sqrt{t} \frac{\mathbb{E} \sup_{s \in [0,t]} |X_s|^m + \mathbb{E} \sup_{s \in [0,t]} |\bar{X}_s|^m + \mathbb{E} \sup_{s \in [0,t]} |\hat{X}_s|^m}{\sqrt{t}} \right]
$$

$$
\leq C \alpha^2 \sqrt{t}.
$$

Let $p > 0$ be the constant in Assumption 2.1 such that $\lim_{t \to \infty} \alpha t^{1/2+2p} = 0$ and for any $\delta \in (0, p)$ define the event $A_{t, \delta} = \{J^{2,1}_t \geq t^{\delta-p}\}$. Then we have for $t$ large enough such that $\alpha t^{1/2+2p} \leq 1$ and

$$
\mathbb{P}(A_{t, \delta}) \leq \frac{\mathbb{E}[J^{2,1}_t]^2}{t^{(\delta-p)}} \leq C \frac{\alpha^2 t^{1/2+2p}}{t^{2\delta}} \leq C \frac{1}{t^{2\delta}}.
$$

The latter implies that

$$
\sum_{n \in \mathbb{N}} \mathbb{P}(A_{2^n, \delta}) < \infty.
$$

Therefore, by the Borel–Cantelli lemma, we have that for every $\delta \in (0, p)$, there is a finite positive random variable $d(\omega)$ and some $n_0 < \infty$ such that for every $n \geq n_0$ one has

$$
J^{2,1}_{2^n} \leq \frac{d(\omega)}{2^{n(p-\delta)}}.
$$

Thus for $t \in [2^n, 2^{n+1})$ and $n \geq n_0$ one has for some finite constant $C < \infty$

$$
J^{2,1}_t \leq C \alpha_{2^{n+1}} \sup_{s \in (0,2^{n+1})} |v(X_s, \bar{X}_s, \hat{X}_s, \vartheta_s)| \leq C \frac{d(\omega)}{2^{(n+1)(p-\delta)}} \leq C \frac{d(\omega)}{t^{p-\delta}},
$$

which derives that for $t \geq 2^{n_0}$, we have with probability one

$$
J^{2,1}_t \leq C \frac{d(\omega)}{t^{p-\delta}} \to 0, \text{ as } t \to \infty.
$$

Next we consider the term

$$
J^{2,2}_{t,0} = \int_0^t \left| \alpha_5 v^2(X_s, \bar{X}_s, \hat{X}_s, \vartheta_s) - 2 \alpha^2 \nabla \vartheta \cdot v^2(X_s, \bar{X}_s, \hat{X}_s, \vartheta_s)(f(\bar{X}_s) - \beta)(\nabla f(X_s)\bar{X}_s)^\top \right| ds
$$
and thus we see that there exists a constant \(0 < C < \infty\) such that
\[
\sup_{t>0} E|J_{t,0}^{2,2}| \leq C \int_0^\infty (|\alpha'_s| + \alpha^2_s) \left(1 + E|X_s|^{m'} + E|X_s|^{m'} + E|X_s|^{m'}\right) ds
\]
\[
\leq C \int_0^\infty (|\alpha'_s| + \alpha^2_s) ds
\]
\[
\leq C,
\]
where in step (a), we use Equation (60) and in step (b), we use Lemma 3.4 and Equation (66). Thus, we know that there is a finite random variable \(J_{\infty,0}^{2,2}\) such that
\[
J_{t,0}^{2,2} \rightarrow J_{\infty,0}^{2,2}, \text{ as } t \rightarrow \infty \text{ with probability one.} \quad (69)
\]

The last term we need to consider is the martingale term
\[
J_{t,0}^{2,3} = \int_0^t \alpha_s \nabla_x v^2(X_s, \bar{X}_s, \bar{X}_s, \theta_s) dW_s + \int_0^t \alpha_s \nabla_{\bar{x}} v^2(X_s, \bar{X}_s, \bar{X}_s, \theta_s) d\bar{W}_s.
\]

Notice that Doob’s inequality and the bounds of Equation (60) (using calculations similar to the ones for the term \(J_{t,0}^{2,2}\)) give us that for some finite constant \(K < \infty\), we have
\[
\sup_{t>0} E|J_{t,0}^{2,3}|^2 \leq K \int_0^\infty \alpha^2_s ds < \infty.
\]

Thus, by Doob’s martingale convergence theorem, there is a square integrable random variable \(J_{\infty,0}^{(3)}\) such that
\[
J_{t,0}^{2,3} \rightarrow J_{\infty,0}^{2,3}, \text{ as } t \rightarrow \infty \text{ both almost surely and in } L^2. \quad (70)
\]

Let us now go back to Equation (64). Using the terms \(J_{t,0}^{2,1}, J_{t,0}^{2,2}, \text{ and } J_{t,0}^{2,3}\), we can write
\[
\left|\Delta_{\sigma_k,\theta_k+\eta_k}^2 J_{t,0}^{2,1} + J_{t,0}^{2,2} + J_{t,0}^{2,3}\right| = \left|\Delta_{\sigma_k,\theta_k+\eta_k}^2 \Psi(\theta)\right|
\]
which together with Equations (68)–(70) prove the statement of the lemma. \(\square\)

Using Equation (B.3) and the dominated convergence theorem, we can establish a bound for the objective function \(J(\theta)\) from Equation (12):
\[
\left|\nabla_{\theta}^2 J(\theta)\right| \leq C \left(\left|E_{Y \sim \pi_{\theta}} f(Y) - \beta\right|^2 + \left|\nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y)\right|^2\right) \leq C, \quad (71)
\]
and therefore the gradient \(\nabla_{\theta} J(\theta)\) is Lipschitz continuous with respect to \(\theta\).
Lemma 3.8. Under Assumption 2.1, choose $\mu > 0$ in Equation (19) such that for the given $\kappa > 0$, one has $3\mu + \frac{\mu}{8\kappa} = \frac{1}{2L_{\nabla \theta}}$, where $L_{\nabla \theta}$ is the Lipschitz constant of $\nabla_{\theta} J(\theta)$ in Equation (12). Then, for $k$ large enough (where $k$ can be random) and $\eta > 0$ small enough (potentially random depending on $k$), $\int_{\tau_k}^{\sigma_k + \eta} \alpha_i ds > \mu$ with probability one. In addition, we also have $\frac{\mu}{2} \leq \int_{\tau_k}^{\sigma_k} \alpha_i ds \leq \mu$ with probability one.

Proof. We use a “proof by contradiction.” Assume that $\int_{\tau_k}^{\sigma_k + \eta} \alpha_i ds \leq \mu$ and let $\delta > 0$ be such that $\delta < \mu/8$. Without loss of generality, we assume that for any $k, \eta$ is small enough such that for any $s \in [\tau_k, \sigma_k + \eta]$, one has $|\nabla_{\theta} J(\theta,s)| \leq 3|\nabla_{\theta} J(\theta_{\tau_k})|$. Combining Equations (17) and (18) yields

$$d\theta_i = -\alpha_i \nabla_{\theta} J(\theta_i) - 2\alpha_i Z_i^1 - 2\alpha_i Z_i^2$$

and thus

$$\left| \theta_{\sigma_k + \eta} - \theta_{\tau_k} \right| \leq \int_{\tau_k}^{\sigma_k + \eta} \alpha_i |\nabla_{\theta} J(\theta_i)| dt + 2 \left| \int_{\tau_k}^{\sigma_k + \eta} \alpha_i Z_i^1 dt \right| + 2 \left| \int_{\tau_k}^{\sigma_k + \eta} \alpha_i Z_i^2 dt \right| \leq 3|\nabla_{\theta} J(\theta_{\tau_k})| \mu + I_1 + I_2.$$  

By Lemmas 3.5 and 3.7, we have that for $k$ large enough,

$$I_1 \leq 2|\Delta^1_{\tau_k,\sigma_k + \eta}| \leq \delta < \mu/16$$

$$I_2 \leq 2|\Delta^2_{\tau_k,\sigma_k + \eta}| \leq \delta < \mu/16.$$  

In addition, we also have by definition that $\frac{\mu}{|\nabla_{\theta} J(\theta_{\tau_k})|} \leq 1$. Combining Equations (73) and (74) yields

$$\left| \theta_{\sigma_k + \eta} - \theta_{\tau_k} \right| \leq |\nabla_{\theta} J(\theta_{\tau_k})| \left(3\mu + \frac{\mu}{8\kappa}\right) = \frac{1}{2L_{\nabla \theta}} |\nabla_{\theta} J(\theta_{\tau_k})|.$$  

This means that

$$\left| \nabla_{\theta} J(\theta_{\sigma_k + \eta}) - \nabla_{\theta} J(\theta_{\tau_k}) \right| \leq L_{\nabla \theta} \left| \theta_{\sigma_k + \eta} - \theta_{\tau_k} \right| \leq \frac{1}{2} |\nabla_{\theta} J(\theta_{\tau_k})|,$$

and thus

$$\frac{1}{2} |\nabla_{\theta} J(\theta_{\tau_k})| \leq |\nabla_{\theta} J(\theta_{\sigma_k + \eta})| \leq 2 |\nabla_{\theta} J(\theta_{\tau_k})|.$$  

However, this produces a contradiction since it implies $\int_{\tau_k}^{\sigma_k + \eta} \alpha_i ds > \mu$; otherwise, from the definition of $\sigma_k$ in Equation (19), we will have $\sigma_k + \eta \in [\tau_k, \sigma_k]$. This concludes the proof of the first part of the lemma.

The proof of the second part of the lemma is straightforward. By its definition in Equation (19), we have that $\int_{\tau_k}^{\sigma_k} \alpha_i ds \leq \mu$. It remains to show that $\int_{\tau_k}^{\sigma_k} \alpha_i ds \geq \frac{\mu}{2}$. We have shown that
\[ \int_{\tau_k}^{\sigma_k + \eta} \alpha_s ds > \mu. \] For \( k \) large enough and \( \eta \) small enough, we can choose that \( \int_{\tau_k}^{\sigma_k + \eta} \alpha_s ds \leq \frac{\mu}{2} \). The conclusion then follows.

**Lemma 3.9.** Under Assumption 2.1, suppose that there exists an infinite number of intervals \( I_k = [\tau_k, \sigma_k] \). Then there is a fixed constant \( \gamma_1 = \gamma_1(\kappa) > 0 \) such that for \( k \) large enough (where \( k \) can be random),

\[ J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) \leq -\gamma_1. \] (75)

**Proof.** By chain rule, we have that

\[ J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) = -\int_{\tau_k}^{\sigma_k} \alpha_\rho \left| \nabla J(\theta_\rho) \right|^2 d\rho - 2 \int_{\tau_k}^{\sigma_k} \alpha_\rho \langle \nabla J(\theta_\rho), Z_1^\rho \rangle d\rho - 2 \int_{\tau_k}^{\sigma_k} \alpha_\rho \langle \nabla J(\theta_\rho), Z_2^\rho \rangle d\rho =: M_{1,k} + M_{2,k} + M_{3,k}. \] (76)

For \( M_{1,k} \), note that for \( \rho \in [\tau_k, \sigma_k] \), we have

\[ \left| \nabla \theta J(\theta_{\tau_k}) \right| \leq \left| \nabla \theta J(\theta_{\rho}) \right| \leq 2 \left| \nabla \theta J(\theta_{\tau_k}) \right|. \] Thus, for sufficiently large \( k \), we have by Lemma 3.8

\[ M_{1,k} \leq -\left| \nabla \theta J(\theta_{\tau_k}) \right|^2 \frac{\mu}{8}. \]

For \( M_{2,k} \) and \( M_{3,k} \), we can use the same method of Poisson equations as in Lemmas 3.5 and 3.7. Define

\[ G^1(x, \bar{x}, \vartheta) = \langle \nabla \theta J(\vartheta), (E_{Y \sim \pi_\vartheta} f(Y) - \beta) \rangle \left( \nabla f(x) \bar{x} - \nabla \theta E_{Y \sim \pi_\vartheta} f(Y) \right)^T \]

\[ G^2(x, \bar{x}, \xi, \vartheta) = \langle \nabla \theta J(\vartheta), (f(\bar{x}) - E_{Y \sim \pi_\vartheta} f(Y)) \rangle \left( \nabla f(x) \bar{x} \right)^T, \] (77)

and use the solution of the corresponding Poisson equations

\[ \mathcal{L}^\vartheta_{x, \bar{x}, \varrho} u^1(x, \bar{x}, \vartheta) = G^1(x, \bar{x}, \vartheta), \]

\[ \mathcal{L}^\vartheta_{x, \bar{x}, \xi, \varrho} u^2(x, \bar{x}, \xi, \vartheta) = G^2(x, \bar{x}, \xi, \vartheta), \] (78)

as in Lemmas 3.5 and 3.7 to prove \( M_{2,k}, M_{3,k} \to 0 \) as \( k \to \infty \) almost surely.

Combining the above results, we obtain that for \( k \) large enough, such that \( |M_{2,k}| + |M_{3,k}| \leq \delta < \frac{\mu}{16} \kappa^2 \)

\[ J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) \leq -\frac{|\nabla J(\theta_{\tau_k})|^2}{8} \mu + \delta \]

\[ \leq -\frac{\mu}{8} \kappa^2 + \frac{\mu}{16} \kappa^2 \]

\[ = -\frac{\mu}{16} \kappa^2. \] (79)
Let \( \gamma_1 = \frac{\mu}{16} \chi^2 \), which concludes the proof of the lemma.

\[ \square \]

**Lemma 3.10.** Under Assumption 2.1, suppose that there exists an infinite number of intervals \( I_k = [\tau_k, \sigma_k) \). Then, there is a fixed constant \( \gamma_2 < \gamma_1 \) such that for \( k \) large enough (where \( k \) can be random),

\[
J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) \leq \gamma_2.
\]

**Proof.** By chain rule, we have

\[
J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) = - \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho |\nabla_\rho J(\theta_\rho)|^2 d\rho + \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\rho J(\theta_\rho), Z_\rho^1 \rangle d\rho + \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\rho J(\theta_\rho), Z_\rho^2 \rangle d\rho
\]

\[
\leq \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\rho J(\theta_\rho), Z_\rho^1 \rangle d\rho + \int_{\sigma_{k-1}}^{\tau_k} \alpha_\rho \langle \nabla_\rho J(\theta_\rho), Z_\rho^2 \rangle d\rho.
\]

(81)

As in the proof of Lemma 3.9, we get that for \( k \) large enough, the right-hand side of the last display can be arbitrarily small, which concludes the proof of the lemma. \[ \square \]

**Proof of Theorem 2.2.** Recalling Equation (19), we know that \( \tau_k \) is the first time \( |\nabla_\theta J(\theta_t)| > \chi \) when \( t > \sigma_{k-1} \). Thus, if for any fixed \( \chi > 0 \), there only exists a finite number of times \( \tau_k \), then there is a finite \( T^* \) such that \( |\nabla_\theta J(\theta_t)| \leq \chi \) for \( t \geq T^* \) and the proof of Equation (2.2) is complete. We now use a "proof by contradiction." Suppose that there are an infinite number of times \( \tau_k \), then by Lemma 3.9 and 3.10, we have for sufficiently large \( k \) (integer \( k \) can be random) that

\[
J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) \leq -\gamma_1
\]

\[
J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) \leq \gamma_2
\]

with \( 0 < \gamma_2 < \gamma_1 \). Choose \( N \) large enough so that the above relations hold simultaneously for \( k \geq N \). Then for all \( n \geq N \),

\[
J(\theta_{\tau_{n+1}}) - J(\theta_{\tau_N}) = \sum_{k=N}^{n} \left[ J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) + J(\theta_{\tau_{k+1}}) - J(\theta_{\sigma_k}) \right]
\]

\[
\leq \sum_{k=N}^{n} (-\gamma_1 + \gamma_2)
\]

\[
< (n - N) \times (-\gamma_1 + \gamma_2).
\]

(82)

Letting \( n \to \infty \), we observe that \( J(\theta_{\tau_n}) \to -\infty \), which is a contradiction, since by definition \( J(\theta_t) \geq 0 \). Thus, there can be at most finitely many \( \tau_k \). Thus, there exists a finite random time \( T \) such that almost surely \( |\nabla_\theta J(\theta_t)| < \chi \) for \( t \geq T \). Since \( \chi \) is arbitrarily chosen, we have proven that \( |\nabla_\theta J(\theta_t)| \to 0 \) as \( t \to \infty \) almost surely. \[ \square \]
4 | NUMERICAL PERFORMANCE OF THE ONLINE ALGORITHM

In this section, we will implement the continuous-time SGD algorithm (4) and evaluate its numerical performance. The algorithm is implemented for a variety of linear and nonlinear models. The algorithm is also implemented for the simultaneous optimization of both the drift and volatility functions, optimizing over a path-dependent SDE, and optimizing over the auto-covariance of an SDE. In our numerical experiments, we found that the performance of the algorithm can depend upon carefully selecting hyperparameters such as the learning rate and mini-batch size. The algorithm with mini-batch size $N$ is

$$
\frac{d\theta_t}{dt} = -2\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} \left( f(X_t^{(i)}) - \beta \right) \right) \cdot \left( \frac{1}{N} \sum_{i=1}^{N} \left( \nabla f(X_t^{(i)}) X_t^{(i)} \right)^T \right),
$$

$$
d\tilde{X}_t^{(i)} = \left( \nabla_x \mu(X_t^{(i)}, \theta_t) \tilde{X}_t^{(i)} + \nabla_\theta \mu(X_t^{(i)}, \theta_t) \right) dt + \left( \nabla_x \sigma(X_t^{(i)}, \theta_t) \tilde{X}_t^{(i)} + \nabla_\theta \sigma(X_t^{(i)}, \theta_t) \right) dW_t^{(i)},
$$

$$
dx_t^{(i)} = \mu(X_t^{(i)}, \theta_t) dt + \sigma(X_t^{(i)}, \theta_t) dW_t^{(i)},
$$

$$
d\tilde{x}_t^{(i)} = \mu(\tilde{X}_t^{(i)}, \theta_t) dt + \sigma(\tilde{X}_t^{(i)}, \theta_t) d\tilde{W}_t^{(i)},
$$

(83)

for $i = 1, 2, \ldots, N$. The notation $(i)$ indicates the $i$th sample in the mini-batch. $\frac{1}{N} \sum_{i=1}^{N} \left( f(X_t^{(i)}) - \beta \right)$ and $\frac{1}{N} \sum_{i=1}^{N} \left( \nabla f(X_t^{(i)}) X_t^{(i)} \right)$ are stochastic estimates of $\mathbb{E}_{Y \sim \pi_\theta}[f(Y) - \beta]$ and $\nabla_\theta (\mathbb{E}_{Y \sim \pi_\theta}[f(X) - \beta])$. A larger mini-batch size reduces the noise in the estimation of the gradient descent direction. The learning rate must decay as $t \to \infty$, but it should not be decreased too rapidly and the initial magnitude should be large enough so that the algorithm converges quickly. In our examples where there is a unique global minimizer, our algorithm will always converge to the optimum if we choose the correct learning rate. For the examples with multiple global minimizers, the algorithm will converge to one of the global minimizers.

**Remark 4.1.** We discuss below some important aspects of the numerical implementation:

(a) Discretization of SDEs: To implement the algorithm (83), we use an Euler scheme with step size $\Delta = 10^{-3} - 10^{-2}$. For example, $X_t^{(i)}$ is simulated as

$$
X_{(n+1)\Delta}^{(i)} = X_{n\Delta}^{(i)} + \left( \nabla_x \mu(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \right) X_{n\Delta}^{(i)} + \nabla_\theta \mu(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \ast \Delta + \left( \nabla_x \sigma(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \right) X_{n\Delta}^{(i)} + \nabla_\theta \sigma(X_{n\Delta}^{(i)}, \theta_{n\Delta}) \ast N(0, 1) \ast \sqrt{\Delta},
$$

(84)

(b) Learning rate and mini-batch size: The learning rate can be chosen to be piecewise constant or gradually decreasing with learning rate schedule

$$
\alpha_t = \frac{C}{1 + t},
$$

where $C$ is also a hyper-parameter needs to be selected. The mini-batch size $N$ that we use is of the order $10^2 - 10^4$. 


Initial values for SDE simulations: In Equation (92), the initial value of the gradient process $\hat{X}_t$ must be zero. The choice of initial points for $X_t, \hat{X}_t$ is flexible. In our experiments, we usually choose $X_0 = \hat{X}_0 = 1$. $\theta_t$ can be randomly initialized or initialized at a deterministic point such as zero.

Objective function: For some simple examples, we can directly calculate the objective function in closed form. For those examples, we directly use that formula to compute the objective function during training. For the more complex examples (with no closed-form formula), we always approximate the objective function $J(\theta)$ using a time-average since, due to the ergodic theorem,

$$
\lim_{t \to \infty} \frac{1}{t} \int_0^t f(X_\theta^s) \, ds = \mathbf{E}_{Y \sim \pi_\theta} f(Y) \quad \text{a.s.} \quad (85)
$$

### 4.1 One-dimensional Ornstein–Uhlenbeck process

We start with a simple case of a one-dimensional Ornstein–Uhlenbeck process $X^\theta_t \in \mathbb{R}$:

$$
dX^\theta_t = (\theta - X^\theta_t) dt + dW_t. \quad (86)
$$

We will use the algorithm (4) to learn the minimizer for

$$
J(\theta) = (\mathbf{E}_{Y \sim \pi_\theta} Y - 2)^2. \quad (87)
$$

Note that in this case, we have the closed-form solution $\pi_\theta \sim N(\theta, \frac{1}{2})$ and thus the global minimizer is $\theta^* = 2$. In Figure 1, several different sample paths generated by the online algorithm are plotted where all trained parameters converges to the global minimizer ($\theta^* = 2$).
Similarly, we use the algorithm (4) to learn the minimizer for

\[ J(\theta) = \left( E_{Y \sim \pi_0} Y^2 - 2 \right)^2. \]

(88)

In this case, the two global minimizers are \( \theta^* = \pm \sqrt{1.5} \). In Figure 2, the parameter trained by the online algorithm converges to a global minimizer. The global minimizer, which the algorithm converges to, depends on the initial value of \( \theta_0 \).

We now consider a more general Ornstein–Uhlenbeck process with parameters \( \theta = (\theta^1, \theta^2) \):

\[
\begin{align*}
    dX^\theta_{\mathbf{1}_i} &= \left( \theta_1 - \theta_2 X^\theta_{\mathbf{1}_i} \right) dt + dW_i, \\
    dX^\theta_{\mathbf{1}_i} &= \left( 1 - \theta_2 X^\theta_{\mathbf{1}_i} \right) dt, \\
    dX^\theta_{\mathbf{2}_i} &= \left( -X^\theta_i - \theta_2 X^\theta_{\mathbf{2}_i} \right) dt, \\
    dX^\theta_i &= \left( \theta_1 - \theta_2 X^\theta_i \right) dt + d\bar{W}_i
\end{align*}
\]

(90)
for $i = 1, 2, \ldots, N$. To make the training more stable and accelerate the convergence rate, we choose the batch size $N = 10,000$. Figures 3 and 4 show the dynamic of the parameters and objective function during training.

4.2 One-dimensional nonlinear process

We now use the online algorithm to optimize over the stationary distribution of a one-dimensional nonlinear process

$$dX_t^\theta = \left(\theta - X_t^\theta - (X_t^\theta)^3\right)dt + dW_t.$$ (91)
We use the algorithm (4) to learn the minimizer of $J(\theta) = (E_{Y \sim \pi_\theta} Y^2 - 2)^2$. The mini-batch algorithm (92) is used:

$$d\theta_t = -4\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} (\bar{X}_t^{(i)})^2 - 2 \right) \cdot \left( \frac{1}{N} \sum_{i=1}^{N} X_t^{(i)} \bar{X}_t^{(i)} \right) dt$$

$$dX_t^{(i)} = \left( \theta_t - X_t^{(i)} - \left( X_t^{(i)} \right)^3 \right) dt + dW_t^{(i)}$$

$$d\bar{X}_t^{(i)} = \left( 1 - \bar{X}_t^{(i)} - 3 \left( X_t^{(i)} \right)^2 \bar{X}_t^{(i)} \right) dt$$

$$d\tilde{X}_t^{(i)} = \left( \theta_t - \tilde{X}_t^{(i)} - \left( \tilde{X}_t^{(i)} \right)^3 \right) dt + d\tilde{W}_t^{(i)}$$

for $i = 1, 2, ..., N$. Figure 5 shows the convergence of the parameter $\theta_t$. In Figure 6, the objective function decays to zero (the global minimum) very quickly.

### 4.3 Optimizing over the drift and volatility coefficients

We now optimize over the drift and volatility functions of the process

$$dX_t^\vartheta = (\mu - X_t^\vartheta) dt + \sigma dW_t$$

(93)
with parameters $\theta = (\mu, \sigma)$. The online algorithm (4) is used to learn the minimizer of $J(\theta) = (\mathbb{E}_{Y \sim \pi_\theta} Y^2 - 2)^2$. The mini-batch algorithm (83) is used:

$$
\frac{d\mu_t}{dt} = -4\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} (\bar{X}_{t}^{(i)})^2 - 2 \right) \cdot \left( \frac{1}{N} \sum_{i=1}^{N} X_{t}^{(i)} \bar{X}_{t}^{(i)} \right) dt
$$

$$
\frac{d\sigma_t}{dt} = -4\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} (X_{t}^{(i)})^2 - 2 \right) \cdot \left( \frac{1}{N} \sum_{i=1}^{N} X_{t}^{(i)} \bar{X}_{t}^{(i)} \right) dt
$$

$$
\frac{dX_{t}^{i}}{dt} = (\mu_t - X_{t}^{(i)}) dt + \sigma_t dW_{t}^{(i)}
$$

$$
\frac{d\bar{X}_{t}^{1,(i)}}{dt} = \left( 1 - \bar{X}_{t}^{1,(i)} \right) dt
$$

$$
\frac{d\bar{X}_{t}^{2,(i)}}{dt} = -\bar{X}_{t}^{2,(i)} dt + dW_{t}^{(i)}
$$

$$
\frac{d\bar{X}_{t}^{(i)}}{dt} = (\mu_t - \bar{X}_{t}^{(i)}) dt + \sigma_t d\bar{W}_{t}^{(i)}
$$

for $i = 1, 2, \ldots, N$. In Figure 7, the trained parameters $\mu_t, \sigma_t$ converge and in Figure 8, the objective function $J(\theta_t) \to 0$ very quickly.

We also implement the online algorithm for the nonlinear process

$$
\frac{dX_{t}^{2}}{dt} = \left( \mu - (X_{t}^{2})^3 \right) dt + \sigma X_{t}^{2} dW_{t},
$$

\[95\]
where $\theta = (\mu, \sigma)$ are the parameters and the objective function is $J(\theta) = (E_{Y \sim \pi_0} Y^2 - 10)^2$. The mini-batch algorithm (83) now becomes

$$d\mu_t = -4\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} (X_t^{(i)})^2 - 2 \right) \cdot \left( \frac{1}{N} \sum_{i=1}^{N} X_t^{(i)} \bar{X}_t^{1,(i)} \right) dt$$
$$d\sigma_t = -4\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} (\bar{X}_t^{(i)})^2 - 2 \right) \cdot \left( \frac{1}{N} \sum_{i=1}^{N} X_t^{(i)} \bar{X}_t^{2,(i)} \right) dt$$
for \(i = 1, 2, \ldots, N\). In Figure 9, the trained parameters \(\mu_t, \sigma_t\) converge and in Figure 10, the objective function \(J(\theta_t) \to 0\) very quickly.

### 4.4 Multidimensional independent Ornstein–Uhlenbeck process

We next consider a simple multidimensional Ornstein–Uhlenbeck process, which consists of \(m\) independent copies of Equation (89). For the parameter \(\theta = (\theta^1, \theta^2) \in R^{2m}\), let the \(m\)-dimensional Ornstein–Uhlenbeck process be

\[
dX_t^\theta = (\theta^1 - \theta^2 \odot X_t^\theta) dt + dW_t,
\]

where \(X_t^\theta \in R^m, W_t \in R^m\), and \(\odot\) is an element-wise product. The objective function is

\[
J(\theta) := \left( \sum_{k=1}^{m} E_{Y \sim \pi_{\theta}} |Y_k|^2 - 2m \right)^2.
\]
The online algorithm (4) is
\[
\begin{align*}
    d\theta_1^t &= -4\alpha_t (|\bar{X}_t|^2 - 2)X_t \odot \bar{X}_1^t dt, \\
    d\theta_2^t &= -4\alpha_t (|X_t|^2 - 2)X_t \odot X_1^t dt, \\
    dX_t &= (\theta_1^t - \theta_2^t \odot X_t) dt + dW^i_t, \\
    d\bar{X}_1^t &= (1 - \theta_2^t \odot \bar{X}_1^t) dt, \\
    d\bar{X}_2^t &= (-X_t - \theta_2^t \odot \bar{X}_2^t) dt, \\
    d\bar{X}_t &= (\theta_1^t - \theta_2^t \odot \bar{X}_t) dt + d\bar{W}_t.
\end{align*}
\]

We implement the algorithm for \( m = 3 \) and \( m = 10 \). In Figures 11 and 12, the objective functions \( J(\theta) \rightarrow 0 \) as \( t \) becomes large.

### 4.5 Multidimensional correlated Ornstein–Uhlenbeck process

For the parameters \( \theta = (\mu, \sigma) \) with \( \mu \in \mathbb{R}^m, \sigma \in \mathbb{R}^{m \times m} \), let the \( m \)-dimensional process \( X_t^\theta \) satisfy
\[
    dX_t^\theta = (\mu - X_t^\theta) dt + \sigma dW_t,
\]
where \( W_t \in \mathbb{R}^m \). Let \( X_t^{\theta, i} \) denote the \( i \)th element of \( X_t^\theta \) and define \( X_t^{\mu, i} \) and \( X_t^{\sigma, i} \) as the Jacobian matrices of \( X_t^\theta \) with respect to \( \mu \) and \( \sigma \):
\[
\begin{align*}
    \bar{X}_t^{\mu} &= \nabla_\mu X_t^\theta \in \mathbb{R}^{m \times m}, \quad \bar{X}_t^{\mu, i} \in \mathbb{R}^m, \\
    \bar{X}_t^{\sigma} &= \nabla_\sigma X_t^\theta \in \mathbb{R}^{m \times m \times m}, \quad \bar{X}_t^{\sigma, i} \in \mathbb{R}^{m \times m}.
\end{align*}
\]
Noting that for $i \in \{1, 2, \ldots, m\}$

$$dX_t^{\beta,i} = \left(\mu_i - X_t^{\beta,i}\right)dt + \sum_j \sigma_{i,j}dW_t^j,$$
now the algorithm (4) becomes

\[
\begin{align*}
    d\mu_i &= -4\alpha_i (|\bar{X}_i|^2 - 2m) \left( \sum_{k=1}^{m} X_i^k \bar{X}_{i}^{\mu,k} \right) dt \\
    d\lambda_i &= -4\alpha_i (|\bar{X}_i|^2 - 2m) \left( \sum_{k=1}^{m} X_i^k \bar{X}_{i}^{\lambda,k} \right) dt \\
    dX_i &= (\mu_i - X_i)dt + \sigma_i dW_i \\
    d\bar{X}_i &= (\mu_i - \bar{X}_i)dt + \sigma_i d\bar{W}_i \\
    d\bar{X}_{i}^{\mu} &= (I_m - \bar{X}_{i}^{\mu})dt \\
    d\bar{X}_{i}^{\sigma,i} &= -\bar{X}_{i}^{\sigma,i} dt + D_i(dW_i), \quad i \in \{1, \ldots, m\}
\end{align*}
\]

where $I_m$ is the $m \times m$ identity matrix and where $D_i(dW_i)$ is a $m \times m$ matrix with all elements equal to 0 except $i$th column being $dW_i$. We examine the algorithm’s performance for dimensions $m = 3, 10$. In Figures 13 and 14, the objective function $J(\theta_i) \to 0$.

### 4.6 Multidimensional nonlinear SDE

In our next example, we optimize over the stationary distribution of a multidimensional nonlinear SDE:

\[
\begin{align*}
    dX_i^\theta &= \left( \theta - \frac{1}{N} \sum_{j=1}^{N} X_i^{\theta,j} - (X_i^{\theta,i})^3 \right) dt + dW_i^i, \quad i = 1, 2, \ldots, N,
\end{align*}
\]
and now \( N \) is the number of agents in the system (103) instead of mini-batch size as before. The objective function is

\[
J(\theta) = \left( \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{Y \sim \pi_\theta} Y_i^2 - 2 \right)^2.
\] (104)

The nonlinear SDE (103) has a mean-field limit as \( N \to \infty \). Thus, for large \( N \), our algorithm could also be used to optimize over the mean-field limit equation (Sznitman, 1991) for Equation (103). The online algorithm for Equation (103) is

\[
\begin{align*}
    d\theta_t &= -4\alpha_t \left( \frac{1}{N} \sum_{i=1}^{N} (X_t^i)^2 - 2 \right) \times \left( \frac{1}{N} \sum_{i=1}^{N} X_t^i X_t^i \right) dt \\
    dX_t^i &= \left( \theta_t - \frac{1}{N} \sum_{j=1}^{N} X_t^j - (X_t^i)^3 \right) dt + dW_t^i \\
    d\bar{X}_t^i &= \left( 1 - \frac{1}{N} \sum_{j=1}^{N} \bar{X}_t^{1,j} - 3(X_t^i)^2 \bar{X}_t^i \right) dt \\
    d\bar{X}_t^i &= \left( \theta_t - \frac{1}{N} \sum_{j=1}^{N} X_t^j - (X_t^i)^3 \right) dt + d\bar{W}_t^i
\end{align*}
\] (105)

for \( i = 1, 2, \ldots, N \). We will select \( N = 1000 \) for our numerical experiment. Therefore, this is an example of high-dimensional SDE model calibration where the dimension of the SDE is \( N = 1000 \). Figures 15 and 16 show the convergence of parameter and objective function.
4.7 Path-dependent SDE

We consider the path-dependent SDE

\[ dX_t^\vartheta = \left( \vartheta - X_t^\vartheta - \frac{1}{t} \int_0^t X_s^\vartheta \, ds \right) dt + dW_t, \]  

(106)
where $X_t^\theta, W_t \in \mathbb{R}$. Although path-dependent SDEs are not directly addressed by this article's convergence theory, this numerical example suggests that the online forward propagation algorithm can also be applied to path-dependent stochastic processes.

For this numerical example, the objective function is

$$J(\theta) = (\mathbb{E}_{Y \sim \pi_\theta} Y - 2)^2.$$  \hfill (107)

The SDE (106) does not fit the problem described in Equations (1) and (2). However, our algorithm still can find the global optimum.

Now the online algorithm (4) is

$$d\theta_t = -4\alpha_t (X_t - 2) X_t dt$$

$$dX_t = \left(\theta_t - X_t - \frac{1}{t} \int_0^t X_s ds\right) dt + dW_t$$

$$d\bar{X}_t = \left(1 - \bar{X}_t - \frac{1}{t} \int_0^t \bar{X}_s ds\right) dt$$

$$d\bar{\bar{X}}_t = \left(\theta_t - \bar{\bar{X}}_t - \frac{1}{t} \int_0^t \bar{\bar{X}}_s ds\right) dt + d\bar{W}_t.$$ \hfill (108)

In Figure 17, the trained parameter converges. The objective function $J(\theta_t)$ is approximated using a time-average. In Figure 18, the objective function $J(\bar{\theta}_t)$ converges to 0 very quickly.
4.8 Optimizing over the auto-covariance of the Ornstein–Uhlenbeck process

As our final numerical example, consider the Ornstein–Uhlenbeck process

\[ dX_t^\theta = (\mu - \lambda X_t^\theta)dt + \sigma dW_t, \] (109)

where \( \theta = (\mu, \lambda, \sigma) \). Define \( \pi_{\theta} \) as the stationary distribution of \( X_t^\theta \) and \( \pi_{\theta,\tau}(dx, dx') \) as the stationary distribution of \( (X_{t-\tau}^\theta, X_t^\theta) \). The objective function is

\[ J(\theta) = \left( \mathbb{E}_{Y \sim \pi_{\theta}} Y - 1 \right)^2 + \left( \mathbb{E}_{Y \sim \pi_{\theta}} Y^2 - 2 \right)^2 + \left( \mathbb{E}_{Y', Y' \sim \pi_{\theta,\tau}} YY' - 1.6 \right)^2, \] (110)

where we will select \( \tau = 0.1 \) for our numerical experiment.

The online algorithm is

\[
\begin{align*}
d\mu_t &= -2\alpha_t [(\tilde{X}_t - 1)\bar{X}_t^1 + 2(\tilde{X}_t^2 - 2)X_t\tilde{X}_t^1 + (\tilde{X}_t - \tau\tilde{X}_t - 1.6)(\bar{X}_t + X_{t-\tau})]dt \\
d\lambda_t &= -2\alpha_t [(\tilde{X}_t - 1)\bar{X}_t^2 + 2(\tilde{X}_t^2 - 2)X_t\tilde{X}_t^2 + (\tilde{X}_t - \tau\tilde{X}_t - 1.6)(\bar{X}_t + X_{t-\tau})]dt \\
d\sigma_t &= -2\alpha_t [(\tilde{X}_t - 1)\bar{X}_t^3 + 2(\tilde{X}_t^2 - 2)X_t\tilde{X}_t^3 + (\tilde{X}_t - \tau\tilde{X}_t - 1.6)(\bar{X}_t + X_{t-\tau})]dt \\
dX_t &= (\mu_t - \lambda_t X_t)dt + \sigma_t dW_t \\
d\tilde{X}_t^1 &= (1 - \lambda_t\tilde{X}_t^1)dt \\
d\tilde{X}_t^2 &= (-X_t - \lambda_t X_t^2)dt \\
d\tilde{X}_t^3 &= -\lambda_t X_t^2 dt + dW_t \\
d\tilde{X}_t &= (\mu_t - \lambda_t\tilde{X}_t)dt + d\tilde{W}_t.
\] (111)
Figures 19–22 display the trained parameters and the objective function. The trained parameters have \~0.1%–0.3% relative error compared to the global minimizers. The objective function \( J(\theta_t) \) is computed from the exact formula:

\[
J(\theta) = \left( \frac{\mu}{\lambda} - 1 \right)^2 + \left( \frac{\mu^2}{\lambda} + \frac{\sigma^2}{2\lambda} - 2 \right)^2 + \left( \frac{\mu}{\lambda} + \frac{\sigma^2 e^{-\lambda t}}{2\lambda} - 1.6 \right)^2.
\] (112)
4.9 Applications to mathematical finance

In this section, we discuss several potential applications of the forward propagation algorithm (4) in mathematical finance. Our algorithm provides a new approach to estimate the parameters in SDE models in mathematical finance and financial econometrics (Aït-Sahalia et al., 2020; Cartea & Jaimungal, 2016; Kitapbayev & Leung, 2018; Lehalle & Neuman, 2019; Leung et al., 2016; Leung & Li, 2015a, 2015b; Zhang et al., 2018), including when the SDE is partially observed. Our algorithm is applicable for the calibration/estimation of SDE model parameters for long time series where ergodicity in the data is expected. In Section 4.10, we discuss parameter estimation in partially observed SDE models (Aït-Sahalia et al., 2020; Sharrock & Kantas, 2022; Surace & Pfister,
In Section 4.11, we discuss the application of our algorithm to solving stochastic optimal control problems for long time horizons where the ergodic framework is suitable; stochastic optimal control is important in many areas of mathematical finance such as optimal order execution and portfolio optimization (Arapostathis et al., 2012; Bardi & Priuli, 2014; Cartea et al., 2015; Hambly et al., 2021; Pham, 2009; Yong & Zhou, 1999). High-dimensional stochastic optimal control problems are computationally intractable for traditional numerical methods. Although the optimal control satisfies a Hamilton–Jacobi–Bellman (HJB) equation, finite difference methods cannot solve high-dimensional PDEs. We demonstrate that our online optimization algorithm can efficiently solve high-dimensional stochastic optimal control problems (in the ergodic setting). In order to evaluate the accuracy of our algorithm for solving stochastic optimal control problems, we implement it for several high-dimensional stochastic linear quadratic regulator (LQR) problems (Bertsekas, 2012; Duncan et al., 1999; Fazel et al., 2018; Hambly et al., 2021; Yong & Zhou, 1999). The LQR problem is selected since a closed-form solution is available (even in high dimensions) to evaluate the accuracy of our algorithm. (However, it should be highlighted that our online optimization algorithm can be used for the stochastic optimal control of any ergodic SDE, including nonlinear SDEs.) The online optimization algorithm learns a parametric control, either a linear function or a neural network (NN), to minimize the objective function. In both the linear and neural network cases, the algorithm can learn the optimal control. The optimal control functions appear in the drift of the SDE. In the case of the neural network optimal control, the SDE is, therefore, a “neural network-SDE.” Neural network SDEs—sometimes referred to as neural-SDEs—are SDEs where the drift and/or volatility of the SDE is a neural network. Neural SDEs have recently become of great interest in mathematical finance (Arribas et al., 2020; Cohen et al., 2023, 2022, 2022b; Gierjatowicz et al., 2020; Ni et al., 2021).

The online optimization algorithm can also be used to solve multi-agent stochastic control problems—for example, mean-field games—which is a widely researched topic in mathematical finance (Bardi & Priuli, 2014; Cao et al., 2022; Cardaliaguet & Mendico, 2021; Carmona et al., 2013; Carmona & Laurière, 2021, 2023) in the ergodic setting. The finite multi-agent stochastic optimal control problem is typically computationally intractable since the corresponding HJB equation is very high-dimensional. It will be an $N \times d$ dimensions PDE, where $N$ is the number of agents and $d$ is the dimension of each agent’s state (i.e., SDE) process. The limit mean-field game, which approximates the finite case, may be computationally tractable to solve. However, if the state space of each agent is high-dimensional (e.g., dimension $d > 4$), the limit mean-field game will also be computationally intractable since it will be a PDE in $d$ dimensions. In addition, the mean-field game limit may not be accurate for the finite-$N$ case if $N$ is not sufficiently large. Therefore, it is of interest to develop new methods for the computational solution of high-dimensional multi-agent stochastic optimal control problems in mathematical finance. As an example, we numerically implement the online optimization model for a simplified version of the multi-agent systemic risk model (Carmona et al., 2013) in Section 4.12. There are $N$ agents where each agent is modeled by an SDE. As $N \to \infty$, the system converges to a mean-field game limit. In the numerical example, we use the online optimization algorithm to solve the high-dimensional stochastic optimal control problem corresponding to a large number of $N$ SDEs ($N = 5000$).

Finally, the online optimization algorithm can be used to train SDE models (including point process models) of limit order books (Bellani et al., 2021; Kumar, 2021; Lu and Abergel, 2018; Morariu-Patrichi & Pakkanen, 2022; Shi & Cartlidge, 2022). Order books involve large numbers
of high-frequency events (∼ 10^5 − 10^6 events per day per stock) and high-dimensional dynamics (many price levels, each with limit order submissions and cancellations, as well as market orders, hidden orders, and transactions). The large amounts of high-frequency high-dimensional data for limit order books makes this a very promising application area for the online forward propagation algorithm, which is able to asymptotically optimize general classes of models over the entire history of the order flow dataset (in contrast to standard methods can typically only optimize over much smaller sub-sequences).

4.10 Optimizing parameters in partially observed SDE models

4.10.1 Two-dimensional Ornstein–Uhlenbeck model

In this section, we focus on the following partially observed two-dimensional Ornstein–Uhlenbeck process (Aït-Sahalia et al., 2020) with parameters \( \vartheta = (\alpha, \sigma_1, \sigma_2) \):

\[
\begin{align*}
    dX_t &= \kappa^1(Y_t - X_t)dt + \sigma^1dW^1_t \\
    dY_t &= \kappa^2(\alpha - Y_t)dt + \sigma^2dW^2_t,
\end{align*}
\]

where the state process \( X_t \) is observable and \( Y_t \) is the latent (unobserved) process. As in Section 4, we can estimate the parameters by calibrating the model to the moments of the stationary distribution. In our numerical example, the objective function is

\[
J(\vartheta) = (E_{Y \sim \pi_\vartheta} Y - 1)^2 + (E_{Y \sim \pi_\vartheta} Y^2 - 2)^2 + (E_{Y \sim \pi_\vartheta} Y^3 - 4)^2.
\]

The algorithm (4) becomes

\[
\begin{align*}
    d\alpha_t &= -\alpha_t \left[ (X_t - 1)X^1_t + 2(X^2_t - 2)X_tX^1_t + 3(X^3_t - 4)X^2_tX^1_t \right] dt \\
    d\sigma^1_t &= -\alpha_t \left[ (X_t - 1)\tilde{X}^1_t + 2(\tilde{X}^2_t - 2)X_t\tilde{X}^1_t + 3(\tilde{X}^3_t - 4)X^2_t\tilde{X}^1_t \right] dt \\
    d\sigma^2_t &= -\alpha_t \left[ (X_t - 1)\tilde{X}^3_t + 2(\tilde{X}^2_t - 2)X_t\tilde{X}^3_t + 3(\tilde{X}^3_t - 4)X^2_t\tilde{X}^3_t \right] dt \\
    dX_t &= \kappa^1(Y_t - X_t)dt + \sigma^1_t dW^1_t \\
    dY_t &= \kappa^2(\alpha - Y_t)dt + \sigma^2_t dW^2_t \\
    d\tilde{X}^1_t &= \kappa^1(Y^1_t - \tilde{X}^1_t) dt \\
    d\tilde{Y}^1_t &= \kappa^2(1 - Y^1_t) dt \\
    d\tilde{X}^2_t &= -\kappa^1\tilde{X}^2_t dt + dW^1_t \\
    d\tilde{X}^3_t &= \kappa^1(Y^3_t - \tilde{X}^3_t) dt \\
    d\tilde{Y}^3_t &= -\kappa^2\tilde{Y}^3_t dt + dW^2_t \\
    d\hat{X}_t &= \kappa^1(Y_t - \hat{X}_t)dt + \sigma^1_t d\hat{W}^1_t \\
    d\hat{Y}_t &= \kappa^2(\alpha_t - \hat{Y}_t)dt + \sigma^2_t d\hat{W}^2_t.
\end{align*}
\]
Figures 23 and 24 display the parameter convergence and the objective function.

4.11 Stochastic optimal control

The online optimization algorithm can be used to solve stochastic optimal control problems, including high-dimensional problems for which traditional numerical methods (e.g., solving the HJB equation with finite difference methods) are computationally expensive or intractable. As a numerical example, we consider the classic LQR problem (Anderson & Moore, 2007; Bertsekas, 2012; Yong & Zhou, 1999), which itself has many financial applications such as optimal execution.
Let \( \{X_t\}_{t \geq 0} \) be the state process that satisfies the SDE
\[
dX_t = (AX_t + BU_t)dt + \sigma dW_t, \tag{116}
\]
where \( X_0 = x_0, X_t \in \mathbb{R}^n \), matrix \( A, \sigma \in \mathbb{R}^{n \times n} \), \( B \in \mathbb{R}^{n \times m} \), \( \{W_t\}_{t \geq 0} \) is an \( \mathbb{R}^n \)-valued standard Wiener process, and \( \{U_t\}_{t \geq 0} \in \mathbb{R}^m \) denotes the control. The objective is to learn a control process \( u_t \) to minimize the following ergodic cost functional for system (116):
\[
J(U) = \lim_{T \to \infty} \frac{1}{T} \int_0^T (X_t^2 QX_t + U_t R U_t) dt,
\tag{117}
\]
where \( Q \) and \( R \) are positive definite matrices. It is well-known that the optimal control is given by Duncan et al. (1999):
\[
U = -R^{-1} B^T K X,
\tag{118}
\]
where \( K \) is the unique solution of the following algebraic Riccati equation (ARE)
\[
A^T K + KA - KB R^{-1} B^T K + Q = 0.
\tag{119}
\]

In order to evaluate the accuracy of our algorithm for solving stochastic optimal control problems, we numerically implement it for several high-dimensional stochastic (LQR) problems. The LQR problem is selected since a closed-form solution is available (even in high dimensions) to evaluate the accuracy of our algorithm. We present a series of numerical examples where the online optimization algorithm learns parametric controls for various LQR problems. The parametric control is either a linear function or a neural network.

### 4.11.1 One-dimensional linear control

As a first step, we implement the online optimization algorithm for the one-dimensional case with a linear control function. For simplicity, we assume that \( A = -1 \), \( B = \sigma = Q = R = 1 \) for Equation (116):
\[
dX_t = (-X_t + \theta X_t)dt + dW_t,
\]
\[
J(\theta) = \lim_{T \to \infty} \frac{1}{T} \int_0^T (X_t^2 + (1 + \theta^2) X_t \bar{X}_t) dt.
\tag{120}
\]

The coupled system (83) becomes
\[
d\theta_t = -\alpha_t \left[ \frac{1}{N} \sum_{i=1}^N \left( 2\theta_i \left(X_i^{(i)}ight)^2 + 2(1 + \theta_i^2)X_i^{(i)} \bar{X}_i^{(i)} \right) \right] dt,
\]
\[
dx_i^{(i)} = (\theta_t - 1)X_i^{(i)} dt + dW_i^{(i)},
\]
\[
dx_i^{(i)} = (X_i^{(i)} + (\theta_t - 1)X_i^{(i)}) dt,
\tag{121}
\]
with $i = 1, 2, \ldots, N$. Solving the ARE (119) yields the optimal control $\theta^* = -0.41421$. Figure 25 shows that the parameter $\theta_t$ trained with the online optimization algorithm converges to $\theta^*$.

### 4.11.2 Multidimensional linear control

We next solve a multidimensional LQR problem with a linear control function. For simplicity, we assume that $m = n$, $A = -I_n$, $B = \sigma = I_n$ in Equation (116) where $I_n$ is $n$ dimensional identity matrix. That is, 

$$
\begin{align*}
    dX_t^\theta &= (-X_t^\theta + \theta X_t^\theta) dt + dW_t, \\
    J(\theta) &= \lim_{T \to \infty} \frac{1}{T} \int_0^T (X_t^\theta)^T (Q + \theta^T R \theta) X_t^\theta dt,
\end{align*}
$$

where $\theta \in \mathbb{R}^{n \times n}$. Let $X_t^{\theta,i}$ denote the $i$th element of $X_t^\theta$ and define 

$$
\begin{align*}
    \dot{X}_t^\theta &= \nabla_\theta X_t^\theta, \quad X_t^{\theta,i} = \nabla_\theta X_t^{\theta,i}, \quad \forall i \in \{1, 2, \ldots, n\}. 
\end{align*}
$$

$\dot{X}_t^\theta$ has dimensions $n \times n \times n$ and $X_t^{\theta,i}$ has dimensions $n \times n$. Note that when we are training over a mini-batch of size $N$, $\dot{X}_t^\theta$ has dimensions $N \times n \times n \times n$.

We first discuss the methods necessary for the computationally efficient simulation of the gradient $\nabla_\theta X_t^\theta$. The state process from Equation (122) satisfies 

$$
\begin{align*}
    dX_t^{\theta,i} &= \left(-X_t^{\theta,i} + \sum_{j=1}^n \theta_{i,j} X_t^{\theta,j}\right) dt + dW_t^i,
\end{align*}
$$
and therefore

\[ dX_t^{\theta,i} = \left( -X_t^{\theta,i} + \sum_{j=1}^{n} \theta_{i,j}X_t^{\theta,j} + D_i \left( X_t^{\theta} \right) \right) dt, \tag{125} \]

where \( D_i(X_t^{\theta}) \) is an \( n \times n \) matrix whose elements are all zeros except for the \( i \)th row, which has values \( X_t^{\theta,i} \). The gradient of the objective function in Equation (122) is

\[ \nabla_{\theta} \left( (X_t^{\theta})^T (Q + \Theta^T R \Theta) X_t^{\theta} \right) = \sum_{i,j} \nabla_{\theta} \left( \delta_{i,j} + \sum_{k=1}^{n} \theta_{k,i} \theta_{k,j} \right) X_t^{\theta,i} X_t^{\theta,j} + 2 \sum_{i,j} \nabla_{\theta} X_t^{\theta,i} \left( q_{i,j} + \Theta^T R \Theta_{j} \right) X_t^{\theta,j}. \tag{126} \]

We now present the method for computationally efficient evaluation of the gradient process \( \dot{X}_t^{\theta} \). For notational simplicity, we only discuss below the case without using a mini-batch. The method can be easily extended to the mini-batch case though. Let \( \odot \) indicate element-wise multiplication with broadcasting (McKinney, 2012). The RHS of Equation (125) can be evaluated using the following operations:

- To vectorize the term \( \sum_{j=1}^{n} \theta_{i,j}X_t^{\theta,j} \), we need to perform an inner-product of the second dimension of the \( n \times n \times 1 \times 1 \) matrix \( \theta \) with the \( 1 \times n \times n \times n \) matrix \( X_t^{\theta} \).
- Note that the final output \( w \) is a tensor with dimensions \( n \times n \times n \).
- To vectorize the term \( D_i(X_t^{\theta}) \), consider the \( n \times n \times n \) tensor \( E \) where \( E_{i,j,:,:} = \delta_{ij} \). Then \( p = E \odot X_t^{\theta} \).
- Add \( w \) and \( p \).

The objective function can be evaluated using a similar method:

- First vectorize the \( (R \Theta)_{i,:,:} + (R \Theta)_{i,:,:} \) to be an \( n \times n \times n \times n \) matrix, which can be achieved by broadcasting, and denote the output as \( D \). Similarly, the matrix multiplication of \( X_t^{\theta,i}X_t^{\theta,j} \) produces a \( n \times n \times 1 \times 1 \), which we denote \( X \).
- Perform an inner-product of the first and second dimension of the \( 1 \times 1 \times n \times n \) matrix \( D \) with the \( n \times n \times 1 \times 1 \) matrix \( X \). Call this output \( z \), which will be a tensor with dimensions \( n \times n \).
- Perform the inner-product of the first dimension of the \( n \times n \times n \) matrix \( X_t^{\theta} \) and \( n \times 1 \times 1 \) matrix \( F \), where \( F_{i,:,:} = \sum_{j}(q_{i,j} + \Theta^T R \Theta_{j})X_t^{\theta,j} \). The output \( q \) is a tensor with dimensions \( n \times n \).
- Add \( z \) and \( q \).

Table 1 presents the numerical results for the online optimization algorithm for learning the optimal control to the LQR problem. The online optimization algorithm performs well even in
high dimensions. Figures 26 and 27 display the maximum and average errors for dimension 5 and 20 during training.

The error metrics in Table 1 are defined as

\[
\text{Ave error} = \frac{\sum_{i,j=1}^{n} \left| \theta_{i,j} - \theta^*_{i,j} \right|}{\sum_{i,j=1}^{n} \left| \theta^*_{i,j} \right|}
\]
Max error = \[ \max_{i,j=\{1,2,\ldots,n\}} \left| \theta_{t,i,j} - \theta^*_i, j \right| + \frac{1}{n^2} \sum_{i,j=1}^{n} \left| \theta^*_i, j \right| \]

Cost error = \[ \frac{|J(\theta_T) - J(\theta^*)|}{|J(\theta^*)|} \], \quad (127) \]

where \( \theta^* \) is the optimal control and \( \theta_t \) is the parameter during training. \( J(\theta_T) \) and \( J(\theta^*) \) denote the objective function \( J(\theta) \) in Equation (122) with the parameters \( \theta_T \) and \( \theta^* \), respectively.

### 4.11.3 One-dimensional neural network control

We will now train a single-layer neural network control using the online optimization algorithm. The state process is

\[
dX_t^q = (-X_t^q + f_{\theta_t}(X_t^q)) dt + dW_t, \quad (128)\]

where the control \( f_{\theta_t}(\cdot) \) is a single-layer neural network

\[
f_{\theta}(x) = \sum_{i=1}^{m} c^i \sigma(w_i x + b_i), \quad (129)\]

with parameters \( \theta = (c^i, w_i, b_i)_{i=1}^m \). The objective function is

\[
J(\theta) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \left( X_t^q \right)^2 + \left( f_{\theta}(X_t^q) \right)^2 dt. \quad (130)\]

Define the gradient of \( X_t \) with respect to the parameters as

\[
\dot{X}_t^w = \nabla_w X_t^q \in \mathbb{R}^m, \quad \dot{X}_t^b = \nabla_b X_t^q \in \mathbb{R}^m, \quad \dot{X}_t^c = \nabla_c X_t^q \in \mathbb{R}^m. \quad (131)\]

The coupled system (4) becomes

\[
dw_t = -\alpha_t \left( 2X_t \dot{X}_t^w + 2f_{\theta_t}(X_t) \left( c_t \odot \sigma'(w_t X_t + b_t) X_t + f'_{\theta_t}(X_t) \dot{X}_t^w \right) \right) dt, \\
db_t = -\alpha_t \left( 2X_t \dot{X}_t^b + 2f_{\theta_t}(X_t) \left( c_t \odot \sigma'(w_t X_t + b_t) X_t + f'_{\theta_t}(X_t) \dot{X}_t^b \right) \right) dt, \\
dc_t = -\alpha_t \left( 2X_t \dot{X}_t^c + 2f_{\theta_t}(X_t) \left( \sigma(w_t X_t + b_t) + f'_{\theta_t}(X_t) \dot{X}_t^c \right) \right) dt, \\
dX_t = (-X_t + f_{\theta_t}(X_t)) dt + dW_t, \\
d\dot{X}_t^w = \left( -\dot{X}_t^w + c_t \odot \sigma' \left( w_t X_t + b_t \right) X_t + f'_{\theta_t}(X_t) \dot{X}_t^w \right) dt, \\
d\dot{X}_t^b = \left( -\dot{X}_t^b + c_t \odot \sigma' \left( w_t X_t + b_t \right) + f'_{\theta_t}(X_t) \dot{X}_t^b \right) dt. \]
The training result for one-dimensional LQR with network network control is presented in Figures 28, 29, and Table 2. The error metrics are defined as:

\[
\begin{align*}
    d\tilde{X}_t^c &= \left(-\tilde{X}_t^c + \sigma (w_tX_t + b_t) + f'_{\tilde{\theta}_t}(X_t)\tilde{X}_t^c\right) dt, \\
    d\tilde{X}_t &= (-\tilde{X}_t + f_{\tilde{\theta}_t}(X_t))dt + d\tilde{W}_t.
\end{align*}
\]
TABLE 2 Training result for NN control.

| Dimension | Ave error | Max error | Cost error |
|-----------|-----------|-----------|------------|
| 1         | 0.1%      | 0.1%      | 0.01%      |
| 5         | 0.6%      | 1%        | 0.02%      |
| 20        | 1%        | 10%       | 0.1%       |

\[
\text{Ave error} = \frac{\sum_{i=1}^{n} \| f_{\theta_t}(X^i) - \theta^* X^i \|}{\sum_{i=1}^{n} \| \theta^* X^i \|}
\]

\[
\text{Max error} = \frac{\max_{i \in \{1, 2, \ldots, n\}} \| f_{\theta_t}(X^i) - \theta^* X^i \|}{\sum_{i=1}^{n} \| \theta^* X^i \|}
\]

\[
\text{Cost error} = \frac{|J(\theta_T) - J(\theta^*)|}{\lambda C^*},
\]

where \( \theta^* \) is the optimal control and \( \theta_t \) is the trained parameter. \( J(\theta_T) \) and \( J(\theta^*) \) denote the objective function \( J(\theta) \) in Equation (130) with the parameters \( \theta_T \) and \( \theta^* \), respectively. The points \( \{X^i\}_{i=1}^{n} \) are uniformly sampled from \([-L, L]\) with \( L \) chosen such that \([-L, L]\) contains the optimally controlled process 99% of the time.

### 4.11.4 Multidimensional neural network control

We now optimize a single-layer neural network control for a high-dimensional state process:

\[
dX_t^\theta = \left( -X_t^\theta + f_\theta(X_t^\theta) \right) dt + dW_t,
\]

\[
J(\theta) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \left( X_t^\theta \right)^T Q X_t^\theta + \left( f_\theta(X_t^\theta) \right)^T R f_\theta(X_t^\theta) dt,
\]

where \( X_t^\theta \in \mathbb{R}^n \) and the single-layer neural network with \( m \) hidden units is

\[
f_\theta(x) = c \sigma(w x + b),
\]

where \( w \in \mathbb{R}^{m \times n} \), \( b \in \mathbb{R}^m \), and \( c \in \mathbb{R}^{n \times m} \). As in Equation (131), define

\[
\begin{align*}
X_t^w &= \nabla_w X_t^\theta \in \mathbb{R}^{n \times m \times n}, & X_t^{w,i} &= \nabla_{w,i} X_t^\theta \in \mathbb{R}^{m \times n}, \\
X_t^b &= \nabla_b X_t^\theta \in \mathbb{R}^{n \times m}, & X_t^{b,i} &= \nabla_{b,i} X_t^\theta \in \mathbb{R}^m, \\
X_t^c &= \nabla_c X_t^\theta \in \mathbb{R}^{n \times m \times m}, & X_t^{c,i} &= \nabla_{c,i} X_t^\theta \in \mathbb{R}^{n \times m}.
\end{align*}
\]

\(^3\)Here the norm \( \| \cdot \| \) denotes the \( L^1 \) norm, that is, for a vector \( Y = (y_1, y_2, \ldots, y_d) \in \mathbb{R}^d \), \( \| Y \| = \sum_{i=1}^{d} |y_i| \).
for \( i = 1, 2, \ldots, n \).

The online algorithm (4) becomes

\[
dw_i = -\alpha_t \left[ \nabla_w \left( f_{\theta_i}(X_t) \right)^T R f_{\theta_i}(X_t) \right] + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left( (X_t)^T Q X_t + f_{\theta_i}(X_t)^T R f_{\theta_i}(X_t) \right) \tilde{X}_{t,i}^w \right] dt
\]

\[
db_i = -\alpha_t \left[ \nabla_b \left( f_{\theta_i}(X_t) \right)^T R f_{\theta_i}(X_t) \right] + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left( (X_t)^T Q X_t + f_{\theta_i}(X_t)^T R f_{\theta_i}(X_t) \right) \tilde{X}_{t,i}^b \right] dt
\]

\[
dc_i = -\alpha_t \left[ \nabla_c \left( f_{\theta_i}(X_t) \right)^T R f_{\theta_i}(X_t) \right] + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left( (X_t)^T Q X_t + f_{\theta_i}(X_t)^T R f_{\theta_i}(X_t) \right) \tilde{X}_{t,i}^c \right] dt
\]

\[
dx_t = (-X_t + f_{\theta_i}(X_t)) dt + dW_t,
\]

\[
d\tilde{X}_{t,i}^w = \left( -\tilde{X}_{t,i}^w + \sum_k c_{t,i,k} \sigma'(w_i X_t + b_i) \left( \sum_\ell \omega_{i,k,\ell} \tilde{X}_{t,\ell}^w \right) + (c_{t,i,:})^T \sigma'(w_i X_t + b_i)(X_t)^T \right) dt
\]

\[
d\tilde{X}_{t,i}^b = \left( -\tilde{X}_{t,i}^b + \sum_k c_{t,i,k} \sigma'(w_i X_t + b_i) \left( \sum_\ell \omega_{i,k,\ell} \tilde{X}_{t,\ell}^b \right) + (c_{t,i,:})^T \sigma'(w_i X_t + b_i) \right) dt
\]

\[
d\tilde{X}_{t,i}^c = \left( -\tilde{X}_{t,i}^c + \sum_k c_{t,i,k} \sigma'(w_i X_t + b_i) \left( \sum_\ell \omega_{i,k,\ell} \tilde{X}_{t,\ell}^c \right) + D_i(\sigma(w_i X_t + b_i)) \right) dt
\]

\[
dX_t = (-X_t + f_{\theta_i}(X_t)) dt + dW_t
\]

(137)

for \( i = 1, 2, \ldots, N \). In Equation (137), \( C_{t,i,:} \in \mathbb{R}^n \) denotes the \( i \)th row of the matrix \( C_t \) and \( D_i(X_t) \) is an \( n \times n \) matrix whose elements are all zeros except for the \( i \)th row, which has the vector value \( \sigma(w_i X_t + b_i) \).

The numerical results for training the neural network SDE control with the online optimization algorithm are presented in Figures 30, 31, and Table 2. In general, the trained neural network control performs well, even in high dimensions.

### 4.12 Applications to multi-agent and mean-field system control

Finally, the online optimization algorithm can be used to solve multi-agent stochastic control problems—for example, mean-field control and mean-field games, which are important topics in mathematical finance (Bardi & Priuli, 2014; Cao et al., 2022; Cardaliaguet & Mendico, 2021; Carmona et al., 2013; Carmona & Laurière, 2021, 2023)—in the ergodic setting. As an example, we numerically implement the online optimization model for a simplified version of the multi-agent systemic risk model (Carmona et al., 2013) in Section 4.12. There are \( N \) agents where each agent is modeled by an SDE. As \( N \to \infty \), the system converges to a mean-field game limit. In the numerical example, we use the online optimization algorithm to solve the high-dimensional stochastic optimal control problem corresponding to a large number of \( N \) SDEs (\( N = 5000 \)).
We consider the following multi-agent control problem, which is a simplified version of the systemic risk model in Carmona et al. (2013):

\[
dX_t^\theta,i = \left[ a \left( \frac{1}{N} \sum_{j=1}^{N} X_t^{\theta,j} - X_t^{\theta,i} \right) + f_\theta \left( X_t^{\theta,i} \right) \right] dt + \sigma dW_t^i
\]  

(138)
for $i = 1, 2, \ldots N$ with the objective function

$$J^N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \lim_{T \to \infty} \frac{1}{T} \int_0^T \left( X_{t,i}^\theta \right)^2 + f^2 \left( X_{t,i}^\theta \right) dt.$$  \hspace{1cm} (139)

This mean-field system has the following mean-field limit:

$$dX_t^\theta = a \left[ \left( E X_t^\theta - X_t^\theta \right) + f( X_t^\theta ) \right] dt + \sigma d W_t$$

$$J(\theta) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \left( X_t^\theta \right)^2 + f^2 \left( X_t^\theta \right) dt.$$ \hspace{1cm} (140)

We describe how the online optimization algorithm can train both linear and neural network controls for this mean-field system. The algorithm (4) to train the linear model becomes

$$d\theta_t = -\alpha_t \left[ \frac{1}{N} \sum_{i=1}^{N} \left( 2\theta_t \left( X_t^i - X_t^j \right)^2 + 2 \left( 1 + \theta_t^2 \right) X_t^i X_t^j \left( X_t^i - X_t^j \right) \right) \right] dt$$

$$dX_t^i = \left[ a \left( \frac{1}{N} \sum_j X_t^j - X_t^i \right) + \theta_t X_t^i \right] dt + dW_t^i$$ \hspace{1cm} (141)

$$d\bar{X}_t^i = \left[ a \left( \frac{1}{N} \sum_j X_t^j - X_t^i \right) + X_t^i + \theta_t X_t^i \right] dt.$$

The training result for the linear control is displayed in Figure 32.
We next train a neural network for the control function $f_\theta(x) = c\sigma(wx + b)$ where $\theta = (c, w, b)$. The online optimization algorithm becomes

$$
dw_t = -\alpha_t \left[ \frac{1}{N} \sum_{i=1}^{N} \left( 2X_t^i \dot{X}_t^i + 2f_{\theta_t}(X_t^i) \left( C_t \odot \sigma' (W_tX_t^i + B_t) X_t^i + f'_{\theta_t}(X_t^i) \dot{X}_t^i \right) \right) \right] dt
$$

$$
db_t = -\alpha_t \left[ \frac{1}{N} \sum_{i=1}^{N} \left( 2X_t^i \dot{X}_t^i + 2f_{\theta_t}(X_t^i) \left( C_t \odot \sigma' (W_tX_t^i + B_t) + f'_{\theta_t}(X_t^i) \dot{X}_t^i \right) \right) \right] dt
$$

$$
dc_t = -\alpha_t \left[ \frac{1}{N} \sum_{i=1}^{N} \left( 2X_t^i \dot{X}_t^i + 2f_{\theta_t}(X_t^i) \left( \sigma (W_tX_t^i + B_t) + f'_{\theta_t}(X_t^i) \dot{X}_t^i \right) \right) \right] dt
$$

$$
dX_t^i = \left[ a \left( \frac{1}{N} \sum_j X_t^j - X_t^i \right) + f_{\theta_t}(X_t^i) \right] dt + dW_t^i
$$

$$
d\tilde{X}_t^{u,i} = \left[ a \left( \frac{1}{N} \sum_j \tilde{X}_t^{u,j} - \tilde{X}_t^{u,i} \right) + C_t \odot \sigma' (W_tX_t^i + B_t) X_t^i + f'_{\theta_t}(X_t^i) \tilde{X}_t^{u,i} \right] dt
$$

$$
d\tilde{X}_t^{b,i} = \left[ a \left( \frac{1}{N} \sum_j \tilde{X}_t^{b,j} - \tilde{X}_t^{b,i} \right) + C_t \odot \sigma' (W_tX_t^i + B_t) + f'_{\theta_t}(X_t^i) \tilde{X}_t^{b,i} \right] dt
$$

$$
d\tilde{X}_t^{c,i} = \left[ a \left( \frac{1}{N} \sum_j \tilde{X}_t^{c,j} - \tilde{X}_t^{c,i} \right) + \sigma (W_tX_t^i + B_t) + f'_{\theta_t}(X_t^i) \tilde{X}_t^{c,i} \right] dt.
$$

(142)

The trained neural network control is also displayed in Figure 32; the controls learned by the linear model and neural network are similar.

### 4.13 Models of order book dynamics

Order books involve large numbers of high-frequency events ($\sim 10^5 - 10^6$ events per day per stock) and high-dimensional dynamics (many price levels, each with limit order submissions and cancellations, as well as market orders, hidden orders, and transactions). Due to the size of the datasets and the high-dimensionality, calibrating simulation models of order book dynamics to data are computationally challenging. Recent examples of such model frameworks for the simulation of the order books include Morariu-Patrichi and Pakkanen (2022), Bellani et al. (2021), Shi and Cartlidge (2022), Lu and Abergel (2018), Kumar (2021), Morariu-Patrichi and Pakkanen (2022), Bellani et al. (2021), Shi and Cartlidge (2022), Lu and Abergel (2018), and Kumar (2021) develop stochastic point process models to model the event-by-event dynamics in order books.

For more complex stochastic models, it is computationally intractable for many traditional calibration methods to optimize over the entire order flow history (even for a few days of events) to estimate the model parameters from the data. The online forward propagation optimization algorithm proposed in this paper provides a tractable computational method to optimize over the entire order flow history. In particular, the online forward propagation optimization algorithm asymptotically minimizes the objective function over the stationary distribution of the entire order
flow process (instead of optimizing over only small subsets of the data, which can lead to a sub-optimal model parameter calibration). In principle, our online optimization algorithm could be used to calibrate a general class of point process models to event-by-event order book data. Such a large-scale data project is outside of the scope of this paper, which is focused on developing a convergence theory. However, in order to demonstrate the applicability of our method to point process models, we present two simple numerical examples below. Synthetic data are simulated from a standard Hawkes process with stochastic intensity

\[
d\lambda_t^\ast = \alpha^\ast (\mu^\ast - \lambda_t^\ast) dt + \kappa^\ast dN_t^\ast,
\]

where \(N_t^\ast\) is the number of events that have occurred by time \(t\). Events arrive with stochastic intensity \(\lambda_t\), i.e., \(\lim_{\Delta \to 0} \frac{P[N_{t+\Delta}^\ast - N_t^\ast = 1 | \mathcal{F}_t]}{\Delta} = \lambda_t^\ast\). For example, \(N_t^\ast\) could be the number of limit orders submitted to the order book by time \(t\). Multidimensional point process models can model the dynamics of the entire order book (e.g., limit order submissions, cancellations, market orders, hidden orders, and transactions) Morariu-Patrichi and Pakkanen (2022), Bellani et al. (2021).

Model parameters for point process models can be calibrated from event data. The data consists of only the observed process \(N_t^\ast\); the stochastic intensity \(\lambda_t^\ast\) is unobserved. Note that Equation (143) is an ergodic process with a stationary distribution. Hawkes process models have been widely used in the financial literature for modeling order book events (for example, see Morariu-Patrichi & Pakkanen, 2022). Using the event data \(N_t^\ast\) simulated from Equation (143), we will calibrate point process models using the online forward propagation optimization algorithm.

First, we consider calibrating a standard Hawkes model using the online optimization algorithm. The model is

\[
d\lambda_t^\vartheta = \alpha (\mu - \lambda_t^\vartheta) dt + \kappa dN_t^\vartheta,
\]

where \(\vartheta = (\alpha, \mu, \kappa)\) are the parameters that must be trained and the time-averaged log-likelihood objective function is

\[
L_T(\vartheta) = -\frac{1}{T} \int_0^T \lambda_t^\vartheta dt + \frac{1}{T} \int_0^T \log(\lambda_t^\vartheta) dN_t^\ast,
\]

where \(\lambda_t^\vartheta\) is the intensity process (144) conditioned on the event observations \((N_t^\ast)_{t \leq T}\), that is, \(d\lambda_t^\vartheta = -\alpha(\mu - \lambda_t^\vartheta) dt + \kappa dN_t^\ast\). Using our online optimization algorithm, we train the parameters \(\vartheta\) to maximize the objective function \(L_T(\vartheta)\). Figure 33 displays the results from the training and demonstrate the numerical convergence of the method. The training converges to a global minimizer; the objective function evaluated at the trained parameters matches the objective function evaluated at the true parameters \(\vartheta^\ast = (\alpha^\ast, \mu^\ast, \kappa^\ast) = (\frac{1}{10}, 1, \frac{1}{10})\).

We now consider a slightly more complex model where the intensity dynamics are given by a neural network. Neural network (or “neural SDEs”) have been widely studied in the financial mathematics literature (Arribas et al., 2020; Cohen et al., 2023, 2022, 2022b; Gierjatowicz et al., 2020; Nie et al., 2021). Neural network Hawkes processes (or “neural Hawkes processes”) have also been recently studied and implemented in a number of papers for modeling order book data (Kumar, 2021; Lu and Abergel, 2018; Shi & Cartlidge, 2022). We consider the following neural
SDE:

\[
d\tilde{\lambda}_t^\theta = f\left(\tilde{\lambda}_t^\theta; \theta\right) \, dt + \kappa d\mathcal{N}_t^\theta,
\]

(146)

where, for this simplified numerical experiment, we set \( \kappa = \kappa^* \) and \( \lambda_t^\theta = |\tilde{\lambda}_t^\theta| + \epsilon \) where \( \epsilon > 0 \). \( f(\lambda; \theta) \) is a single-layer neural network with 25 hidden units. The neural network parameters \( \theta \) are trained with the online forward propagation optimization algorithm:

\[
d\tilde{\lambda}_t = \left(\frac{\partial f}{\partial \lambda}(\tilde{\lambda}_t; \theta_t)\tilde{\lambda}_t + \frac{\partial f}{\partial \theta}(\tilde{\lambda}_t; \theta_t)\right) dt,
\]

\[
d\tilde{\lambda}_t = f(\lambda_t; \theta_t)dt + \kappa d\mathcal{N}_t^\theta,
\]

\[
d\theta_t = \alpha_t \left( - \frac{\partial \lambda_t}{\partial \tilde{\lambda}_t}d\tilde{\lambda}_t + (\lambda_t)^{-1} \frac{\partial \lambda_t}{\partial \tilde{\lambda}_t} \mathcal{N}_t^\theta \right),
\]

(147)

where \( \lambda_t = |\tilde{\lambda}_t| + \epsilon \) and \( \alpha_t \) is the learning rate. The data \( \mathcal{N}_t^\theta \), which the model (146) is trained on, is generated using Equation (143) with the “true parameters” \( \theta^* = (\frac{1}{10}, 1, \frac{1}{10}) \). The training and out-of-sample test results are displayed in Figure 34. The plots display the value of the objective function (145) evaluated using the “true” process (143) with the true parameters \( \theta^* \) (which is the global minimum) as compared to the value of the objective function (145) for the trained model (146). The neural network point process model (146), trained with the online forward propagation algorithm, is able to achieve a nearly identical value for the objective function as the exact global minimizer (with \( \sim 10^{-4} \) relative error), indicating that the trained model converges to a global minimizer.

We conclude by highlighting that—although outside of the scope of this paper—a more general multidimensional model for the entire order book (see Morariu-Patrichi & Pakkanen, 2022) could also be calibrated to real order book data using the online forward propagation algorithm. General classes of multidimensional neural SDE models can be optimized using our method. For example, “recurrent neural SDEs,” where the dynamics (144) depend upon the evolution of a “hidden” neural SDE, can also be calibrated using the online forward propagation method, such as

\[
d\tilde{\lambda}_t^\theta = f\left(\tilde{\lambda}_t^\theta, S_t^\theta; \theta\right) \, dt + \kappa (\tilde{\lambda}_t^\theta, S_t^\theta; \theta) \, d\mathcal{N}_t^\theta,
\]
where $f, g, h, \lambda, \xi$ are neural networks with collective parameters $\theta$ and where $\lambda^\theta_i, N^\theta_i$, and $S^\theta_i$ can be multidimensional. Recurrent neural networks Hawkes models for order books have been investigated in Kumar (2021) and Shi and Cartlidge (2022). Recurrent neural network Hawkes processes have recently received significant interest in the broader machine learning community (Mei & Eisner, 2017). General classes of continuous-time recurrent network SDEs have also been proposed in Munikoti et al. (2023). A more general class of continuous-time recurrent network point processes has also been developed in Chen et al. (2020); Equation (148) is an example from the general framework in Chen et al. (2020). The unique capability provided by the algorithm is to asymptotically optimize such models over the entire history of the order flow dataset, while standard methods can typically only optimize over much smaller subsequences.

5 | CONCLUSION

In this paper, we proposed a new online algorithm for computationally efficient optimization over the stationary distribution of ergodic SDEs. In particular, the online forward propagation algorithm can optimize over parameterized SDEs in order to minimize the distance between their stationary distribution and target statistics. By proving bounds for a new class of Poisson PDEs, we can analyze the parameters’ fluctuations during training and rigorously prove convergence to a stationary point for linear SDE models. We also study the numerical performance of our algorithm for nonlinear examples. In the nonlinear cases, which we present in this paper, the algorithm performs well and the parameters converge to a minimizer.

Our algorithm can be used for applications where optimizing over the stationary distribution of an SDE model is of interest. In many applications, the stationary distribution $\pi^\theta$ is unknown and the dimension of the stochastic process may be large. The online algorithm developed in this paper is well-suited for such problems.

Finally, there are several future research directions, which should be explored. First, a convergence analysis for nonlinear SDEs would be an important next step. The focus of our paper is a convergence analysis for linear SDEs; this required addressing several nontrivial mathematical challenges, in particular, the development and rigorous analysis of a new class of Poisson PDEs. Our
results in this paper provide the building blocks for a future nonlinear analysis. The convergence of our online algorithm for discrete-time stochastic processes would also be interesting to study.

**ACKNOWLEDGMENTS**

This research has been supported by the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modeling, and Simulation (EP/S023925/1).

**DATA AVAILABILITY STATEMENT**

This article does not use any datasets.

**ORCID**

Justin Sirignano https://orcid.org/0000-0002-0971-1349

**REFERENCES**

Aït-Sahalia, Y., Li, C., & Li, C. X. (2020). Maximum likelihood estimation of latent Markov models using closed-form approximations. *Journal of Econometrics*, 105008.

Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3, 5–40.

Anderson, B. D., & Moore, J. B. (2007). *Optimal control: Linear quadratic methods*. Courier Corporation.

Annunziato, M., & Borzì, A. (2013). A Fokker–Planck control framework for multidimensional stochastic processes. *Journal of Computational and Applied Mathematics*, 237(1), 487–507.

Arapostathis, A., Borkar, V. S., & Ghosh, M. K. (2012). *Ergodic control of diffusion processes* (Vol. 143). Cambridge University Press.

Arribas, I. P., Salvi, C., & Szpruch, L. (2020). Sig-SDEs model for quantitative finance. In *Proceedings of the First ACM International Conference on AI in Finance*.

Bardi, M., & Priuli, F. S. (2014). Linear-quadratic n-person and mean-field games with ergodic cost. *SIAM Journal on Control and Optimization*, 52(5), 3022–3052.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, 9(1), 69–107.

Bellani, C., Brigo, D., & Pakkanen, M. (2021). Non-average price impact in order-driven markets. *arXiv preprint arXiv:2110.00771*.

Benveniste, A., Métivier, M., & Priouret, P. (2012). *Adaptive algorithms and stochastic approximations* (Vol. 22). Springer Science & Business Media.

Benzing, F., Gauy, M., Mujika, A., Martinsson, A., & Steger, A. (2019). Optimal kronecker-sum approximation of real time recurrent learning. In *Proceedings of the 36th International Conference on Machine Learning*.

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I* (Vol. 1). Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3), 627–642.

Bhudisaksang, T., & Cartea, Á. (2021). Online drift estimation for jump-diffusion processes. *Bernoulli*, 27(4), 2494–2518.

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2), 223–311.

Butt, M. (2022). Numerical solution to 3D bilinear Fokker–Planck control problem. *International Journal of Computer Mathematics*, 99(12), 2466–2481.

Cao, H., Dianetti, J., & Ferrari, G. (2022). Stationary discounted and ergodic mean field games with singular controls. *Mathematics of Operations Research*, 48(4), 1871–1898.

Cardaliaguet, P., & Mendico, C. (2021). Ergodic behavior of control and mean field games problems depending on acceleration. *Nonlinear Analysis*, 203, 112185.

Carmona, R., Fouque, J.-P., & Sun, L.-H. (2015). Mean field games and systemic risk. *Communications in Mathematical Sciences*, 13(4), 911–933.

Carmona, R., & Laurière, M. (2021). Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games I: The ergodic case. *SIAM Journal on Numerical Analysis*, 59(3), 1455–1485.
Carmona, R., & Laurière, M. (2023). Deep learning for mean field games and mean field control with applications to finance. *Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices*, 369.

Cartea, A., & Jaimungal, S. (2016). Incorporating order-flow into optimal execution. *Mathematics and Financial Economics*, 10, 339–364.

Cartea, A., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and high-frequency trading*. Cambridge University Press.

Chen, R. T. Q., Amos, B., & Nickel, M. (2020). Neural spatio-temporal point processes. *International Conference on Learning Representations*.

Christoffersen, P., Heston, S., & Jacobs, K. (2009). The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. *Management Science*, 55(12), 1914–1932.

Cohen, S. N., Reisinger, C., & Wang, S. (2022). Hedging option books using neural-SDE market models. *Applied Mathematical Finance*, 1–36.

Cohen, S. N., Reisinger, C., & Wang, S. (2023). Arbitrage-free neural-SDE market models. *Applied Mathematical Finance*, 1–46.

Cohen, S. N., Reisinger, C., & Wang, S. (2022b). Hedging option books using neural-SDE market models. *arXiv preprint arXiv:2205.15991*.

Collin-Dufresne, P., & Goldstein, R. S. (2002). Do bonds span the fixed income markets? Theory and evidence for unspanned stochastic volatility. *The Journal of Finance*, 57(4), 1685–1730.

Collin-Dufresne, P., Goldstein, R. S., & Jones, C. S. (2009). Can interest rate volatility be extracted from the cross section of bond yields? *Journal of Financial Economics*, 94(1), 47–66.

Creal, D. D., & Wu, J. C. (2015). Estimation of affine term structure models with spanned or unspanned stochastic volatility. *Journal of Econometrics*, 185(1), 60–81.

Duffie, D., Pan, J., & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6), 1343–1376.

Duncan, T. E., Guo, L., & Pasik-Duncan, B. (1999). Adaptive continuous-time linear quadratic gaussian control. *IEEE Transactions on automatic control*, 44(9), 1653–1662.

Fazel, M., Ge, R., Kakade, S., & Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning* (pp. 1467–1476). PMLR.

Fleig, A., & Guglielmi, R. (2017). Optimal control of the Fokker–Planck equation with space-dependent controls. *Journal of Optimization Theory and Applications*, 174, 408–427.

Folland, G. B. (1999). *Real analysis: Modern techniques and their applications* (Vol. 40). John Wiley & Sons.

Gierzatowicz, P., Sabate-Vidales, M., Siska, D., Szpruch, L., & Zuric, Z. (2020). Robust pricing and hedging via neural SDEs. *arXiv preprint arXiv:2007.04154*.

Gilbarg, D., & Trudinger, N. S. (2015). *Elliptic partial differential equations of second order* (Vol. 224). Springer.

Giles, M., & Fang, W. (2020). Adaptive Euler-Maruyama method for SDEs with non-globally Lipschitz drift. *Annals of Applied Probability*, 30(2).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hambly, B., Xu, R., & Yang, H. (2021). Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5), 3359–3391.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2), 327–343.

Kaltenbacher, B., & Pedretsch, B. (2018). Parameter estimation in SDEs via the Fokker–Planck equation: Likelihood function and adjoint based gradient computation. *Journal of Mathematical Analysis and Applications*, 465(2), 872–884.

Kitapbayev, Y., & Leung, T. (2018). Mean reversion trading with sequential deadlines and transaction costs. *International Journal of Theoretical and Applied Finance*, 21(01), 1850004.

Kumar, P. (2021). Deep Hawkes process for high-frequency market making. *arXiv preprint arXiv:2109.15110*.

Kushner, H., & Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications* (Vol. 35). Springer Science & Business Media.

Lehalle, C.-A., & Neuman, E. (2019). Incorporating signals into optimal trading. *Finance and Stochastics*, 23, 275–311.

Leung, T., Li, J., Li, X., & Wang, Z. (2016). Speculative futures trading under mean reversion. *Asia-Pacific Financial Markets*, 23, 281–304.
Leung, T., & Li, X. (2015a). Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical and Applied Finance, 18*(03), 1550020.

Leung, T. S.-t., & Li, X. (2015b). Optimal mean reversion trading: Mathematical analysis and practical applications (Vol. 1). World Scientific.

Lu, X., & Abergel, F. (2018). High-dimensional hawkes processes for limit order books: Modelling, empirical analysis and numerical calibration. *Quantitative Finance, 18*(2), 249–264.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O’Reilly Media, Inc.

Mei, H., & Eisner, J. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems* (Vol. 30).

Menick, J., Elsen, E., Evci, U., Osindero, S., Simonyan, K., & Graves, A. (2018). A practical sparse approximation for real time recurrent learning. *arXiv:2006.07232*.

Morariu-Patrichi, M., & Pakkanen, M. (2022). State-dependent Hawkes processes and their application to limit order book modelling. *Quantitative Finance, 22*(3), 563–583.

Munikoti, S., Dahale, S., & Natarajan, B. (2023). A general framework for uncertainty quantification via neural SDE-RNN. *arXiv preprint arXiv:2306.01189*.

Ni, H., Szpruch, L., Vidalas, M., Xiao, B., Wiese, M., & Liao, S. (2021). Sig-Wasserstein GANs for time series generation. In *Proceedings of the Second ACM International Conference on AI in Finance*.

Pardoux, E., & Veretennikov, A. Y. (2001). On the Poisson equation and diffusion approximation. I. *Annals of Probability, 29*, 1061–1085.

Pardoux, E., & Veretennikov, A. Y. (2003). On Poisson equation and diffusion approximation 2. *The Annals of Probability, 31*(3), 1166–1192.

Pavliotis, G. A. (2014). *Stochastic processes and applications*. Springer.

Pham, H. (2009). *Continuous-time stochastic control and optimization with financial applications* (Vol. 61). Springer Science & Business Media.

Röckner, M., Sun, X., & Xie, Y. (2021). Strong convergence order for slow–fast Mckean–Vlasov stochastic differential equations. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* (Vol. 57, pp. 547–576). Institut Henri Poincaré.

Sharrock, L., & Kantas, N. (2022). Joint online parameter estimation and optimal sensor placement for the partially observed stochastic advection-diffusion equation. *SIAM/ASA Journal on Uncertainty Quantification, 10*(1), 55–95.

Sharrock, L., & Kantas, N. (2023). Two-timescale stochastic gradient descent in continuous time with applications to joint online parameter estimation and optimal sensor placement. *Bernoulli, 29*(2), 1137–1165.

Sharrock, L., Kantas, N., Parpas, P., & Pavliotis, G. A. (2021). Online parameter estimation for the McKean-Vlasov stochastic differential equation. *Stochastic Processes and their Applications, 162*, 481–546.

Shi, Z., & Cartlidge, J. (2022). State dependent parallel neural Hawkes process for limit order book event stream prediction and simulation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1607–1615).

Sirignano, J., & Spiliopoulos, K. (2017). Stochastic gradient descent in continuous time. *SIAM Journal on Financial Mathematics, 8*(1), 933–961.

Sirignano, J., & Spiliopoulos, K. (2020). Stochastic gradient descent in continuous time: A central limit theorem. *Stochastic Systems, 10*(2), 124–151.

Surace, S. C., & Pfister, J.-P. (2018). Online maximum-likelihood estimation of the parameters of partially observed diffusion processes. *IEEE Transactions on Automatic Control, 64*(7), 2814–2829.

Sznitman, A.-S. (1991). Topics in propagation of chaos. *Ecole d’été de probabilités de Saint-Flour XIX—1989, 1464, 165–251*.

Wang, Z., & Sirignano, J. (2022). A forward propagation algorithm for online optimization of nonlinear stochastic differential equations. *arXiv preprint arXiv:2207.04496*.

Williams, R., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation, 1*(2), 270–280.

Yong, J., & Zhou, X. Y. (1999). *Stochastic controls: Hamiltonian systems and HJB equations* (Vol. 43). Springer Science & Business Media.
How to cite this article: Wang, Z., & Sirignano, J. (2024). Continuous-time stochastic gradient descent for optimizing over the stationary distribution of stochastic differential equations. *Mathematical Finance*, 34, 348–424. https://doi.org/10.1111/mafi.12422

APPENDIX A: PROOF OF PROPOSITION 3.1

We first present a useful lemma before proving Proposition 3.1. The bound (A.1) will be frequently used in the proof of Proposition 3.1.

**Lemma A.1.** For any \( m', k \in \mathbb{R}_+ \), there exist constants \( C, m > 0 \) such that for any \( x, x' \in \mathbb{R}^d \),

\[
e^{-|x'-x|^2} \cdot |x' - x|^k \leq C \frac{1 + |x|^m}{1 + |x'|^{m'}}. \tag{A.1}
\]

**Proof.** For any fixed \( x \in \mathbb{R}^d \), when \( \frac{|x'|}{2} \geq |x| \), we have

\[
|x' - x| \geq |x'| - |x| \geq \frac{|x'|}{2},
\]

\[
|x' - x| \leq |x'| + |x| \leq \frac{3|x'|}{2}. \tag{A.2}
\]

Therefore, we have that for any \( m', k > 0 \), there exists a constant \( C_1 > 0 \) such that

\[
e^{-|x'-x|^2} \cdot |x' - x|^k \leq e^{-\frac{|x'|^2}{4}} \left( \frac{3}{2} |x'| \right)^k \leq C_1 \frac{1}{1 + |x'|^{m'}}. \tag{A.3}
\]

where the first inequality is due (A.2) and step (a) uses the fact that

\[
\lim_{s \to +\infty} \frac{s^m}{e^s} = 0, \quad \forall m > 0. \tag{A.4}
\]

When \( \frac{|x'|}{2} < |x| \), for any \( m', k > 0 \), there exist constants \( C_2, m > 0 \) such that

\[
e^{-|x'-x|^2} \cdot |x' - x|^k \leq (3|x|)^k \frac{1 + 2|x'|^m}{1 + |x'|^{m'}} \leq C_2 \frac{1 + |x|^m}{1 + |x'|^{m'}}. \tag{A.5}
\]

Let us now choose \( C = C_1 + C_2 \) and then Equation (A.1) holds. \( \square \)

**Proof of Proposition 3.1.** The proof for the convergence results leverages the closed-form formula for the distribution. Let,

\[
f(t, x, \theta) = e^{-h(\theta)t}x + h(\theta)^{-1}(I_d - e^{-h(\theta)t})g(\theta), \quad \Sigma_t(\theta) = \sigma^2(2h(\theta))^{-1}(I_d - e^{-2h(\theta)t}), \tag{A.6}
\]
and from Equation (11), we know that
\[ X_t^\theta \sim N(f(t, x, \theta), \Sigma_t(\theta)). \]  
(A.7)

Thus, the stationary distribution for \( X_t^\theta \) is \( N(h^{-1}(\theta)g(\theta), \sigma^2(2h(\theta))^{-1}) \). Since \( h(\theta) \) is positive definite, there exists orthogonal matrix \( Q(\theta) \) such that
\[ h(\theta) = Q(\theta)^T \Lambda(\theta) Q(\theta) \]
where \( \Lambda(\theta) = \text{diag}(\lambda_1(\theta), ..., \lambda_d(\theta)) \) is a diagonal and all its eigenvalues are positive. Thus for \( t > 0 \)
\[ \Sigma_t(\theta) = \frac{\sigma^2}{2} Q(\theta)^T \Lambda^{-1}(\theta) (I_d - e^{-2\Lambda(\theta)t}) Q(\theta), \]  
(A.8)

and the eigenvalues of \( \Sigma_t(\theta) \) are \( (\frac{\sigma^2(1-e^{-2\lambda_1(\theta)t})}{2\lambda_1(\theta)}, ..., \frac{\sigma^2(1-e^{-2\lambda_d(\theta)t})}{2\lambda_d(\theta)}) \). Then, we know that the covariance matrix \( \Sigma_t(\theta) \) is also positive definite for any \( t > 0 \) and the density is
\[ p_t(x, x', \theta) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_t(\theta)|}} \exp \left\{ -\frac{1}{2} \left( x' - f(t, x, \theta) \right)^T \Sigma_t^{-1}(\theta) \left( x' - f(t, x, \theta) \right) \right\}, \quad t > 0 \]  
(A.9)

Proof of (i). Recall that (by assumption) \( h(\theta) \) is uniformly positive definite and thus
\[ p_\infty(x', \theta) = C \sqrt{|h(\theta)|} \exp \left\{ -c |x' - h(\theta)^{-1}g(\theta)|^2 \right\} \leq C \]  
\[ \frac{1}{1 + |x'|^m} \]
(A.10)

where step (a) uses the bounds for \( g, h \) in Assumption 2.1 and Equation (A.4). Due to Equation (A.9), we have for any \( k \in \{1, 2, ..., \ell\} \) that
\[
\left| \frac{\partial}{\partial \theta_k} p_\infty(x', \theta) \right| \\
\leq C \left( \frac{\partial}{\partial \theta_k} \sqrt{|h(\theta)|} \right) \cdot \exp \left\{ -c |x' - h(\theta)^{-1}g(\theta)|^2 \right\} + C \sqrt{|h(\theta)|} \exp \left\{ -c |x' - h(\theta)^{-1}g(\theta)|^2 \right\} \\
\left| \left( x' - h(\theta)^{-1}g(\theta) \right)^T \frac{\partial h(\theta)}{\partial \theta_k} (x' - h(\theta)^{-1}g(\theta)) \right| + 2 \left( \frac{\partial h(\theta)^{-1}g(\theta)}{\partial \theta_k} \right)^T h(\theta) (x' - h(\theta)^{-1}g(\theta)) \\
\leq C \exp \left\{ -c |x' - h(\theta)^{-1}g(\theta)|^2 \right\} + C \exp \left\{ -c |x' - h(\theta)^{-1}g(\theta)|^2 \right\} \left| \left( x' - h(\theta)^{-1}g(\theta) \right)^2 + |x' - h(\theta)^{-1}g(\theta)| \right| \\
\leq \frac{C}{1 + |x'|^m}, \]
(A.11)
where step (a) is by the boundedness of \(g(\theta), \frac{\partial g(\theta)}{\partial \theta_k}, h(\theta), \frac{\partial h(\theta)}{\partial \theta_k}\) and since \(h(\theta)\) is positive definite due to Assumption 2.1. Step (b) is due to Equation (A.1) with \(x = h(\theta)^{-1}g(\theta)\) and Equation (A.4). Using the same method as in Equation (A.11), we can obtain the bound for \(\nabla^2 \psi \|\psi\|_\infty(x, \theta)\).

**Proof of (ii) and (iii).** We now prove Equation (26). First let

\[
X := x' - f(t, x, \theta), \quad Y := x' - h(\theta)^{-1}g(\theta),
\]

and then since \(h\) is uniformly positive definite:

\[
|X - Y| = \left| e^{-h(\theta)t}x - e^{-2h(\theta)t}h(\theta)^{-1}g(\theta) \right| \leq Ce^{-ct}(1 + |x|). \tag{A.12}
\]

We will use the following decomposition:

\[
|p_\psi(x, x', \theta) - p_\psi(\xi', \theta)| \\
\leq C \left( \frac{1}{\sqrt{|(I_d - e^{-2h(\theta)t})\psi|}} - 1 \right) + C \left| \exp \left\{ -X^T \frac{h(\theta)}{\sigma^2} (I_d - e^{-2h(\theta)t})^{-1} X \right\} - \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} (I_d - e^{-2h(\theta)t})^{-1} Y \right\} \right| \\
+ \left| \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} (I_d - e^{-2h(\theta)t})^{-1} Y \right\} - \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} Y \right\} \right| \\
= : I_1 + I_2 + I_3 \tag{A.13}
\]

For \(I_1\), note that when \(t > 1\),

\[
\frac{1}{\sqrt{|(I_d - e^{-2h(\theta)t})\psi|}} - 1 = \frac{1}{\sqrt{\prod_{k=1}^{d} (1 - e^{-2\lambda_k(\theta)t})}} - 1 \leq C \left[ 1 - \prod_{k=1}^{d} (1 - e^{-2\lambda_k(\theta)t}) \right] \leq Ce^{-2\lambda_1(\theta)t} \leq Ce^{-2ct} \tag{A.14}
\]

where \(\lambda_1(\theta) \leq \lambda_2(\theta) \leq \cdots \leq \lambda_d(\theta)\) are the eigenvalues of the matrix \(h(\theta)\). For \(I_3\), similar to Equation (A.8), we know that the eigenvalues of \(h(\theta)((I_d - e^{-2h(\theta)t})^{-1} - I_d)\) are \(\frac{\lambda_i(\theta)e^{-2\lambda_i(\theta)t}}{1 - e^{-2\lambda_i(\theta)t}}, i = 1, \ldots, d\), which implies that \(h(\theta)((I_d - e^{-2h(\theta)t})^{-1} - I_d)\) is also a positive definite matrix. When \(t > 1\), since \(h(\theta)\) is uniformly positive definite, the eigenvalues will have a uniform upper bound:

\[
\frac{\lambda_i(\theta)e^{-2\lambda_i(\theta)t}}{1 - e^{-2\lambda_i(\theta)t}} \leq \frac{C}{1 - e^{-c}} \leq C, \quad t > 1, \; \forall i \in \{1, \ldots, d\}. \tag{A.15}
\]

Thus for any \(m', k > 0\), there exists a constant \(C > 0\) such that when \(t > 1\),

\[
\left| \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} (I_d - e^{-2h(\theta)t})^{-1} Y \right\} - \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} Y \right\} \right| \\
= \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} Y \right\} \left| \exp \left\{ -Y^T \frac{h(\theta)}{\sigma^2} ((I_d - e^{-2h(\theta)t})^{-1} - I_d) Y \right\} - 1 \right|
\]
\[(a) \quad C \exp \left\{ -Y^\top \frac{h(\theta)}{\sigma^2} Y \right\} \left| Y^\top \frac{h(\theta)}{\sigma^2} \left( (I_d - e^{-2h(\theta)t})^{-1} - I_d \right) Y \right| \]
\[(b) \quad C \exp \left\{ -Y^\top \frac{h(\theta)}{\sigma^2} Y \right\} |Y|^2 \cdot \lambda_{\text{max}} \left( h(\theta) \left( (I_d - e^{-2h(\theta)t})^{-1} - I_d \right) \right) \]
\[(c) \quad C \exp \left\{ -c |x' - h(\theta)^{-1} g(\theta)|^2 \right\} \cdot |x' - h(\theta)^{-1} g(\theta)|^2 \]
\[(d) \quad C e^{-ct} \frac{1}{1 + |x'|^m (1 + t)^k}, \quad (A.16)\]

where step (a) is by the positive definiteness of \( \frac{h(\theta)}{\sigma^2} \left( (I_d - e^{-2h(\theta)t})^{-1} - I_d \right) \), which means
\[Y^\top \frac{h(\theta)}{\sigma^2} \left( (I_d - e^{-2h(\theta)t})^{-1} - I_d \right) Y \geq 0,\]
and the fact \( 0 \leq 1 - e^{-s} \leq s, \forall s \geq 0 \). In step (b), \( \lambda_{\text{max}} \) denotes the largest eigenvalue and step (c) uses Equation (A.15). Step (d) follows from Equation (A.1) with \( x = h(\theta)^{-1} g(\theta) \) and the boundedness of \( g, h \).

For \( I_2 \), define the function on \( F_t : \mathbb{R}^d \to \mathbb{R} \) for \( t > 0 \)
\[F_t(x) := \exp \left\{ -x^\top \frac{h(\theta)}{\sigma^2} \left( (I_d - e^{-2h(\theta)t})^{-1} \right) x \right\}.\]

By mean value theorem,
\[F_t(x) - F_t(y) = \nabla F_t(x_0)^\top (x - y) = -\frac{2h(\theta)}{\sigma^2} \left( (I_d - e^{-2h(\theta)t})^{-1} \right) F_t(x_0) x_0^\top (x - y), \quad (A.17)\]
where \( x_0 = t_0 x + (1 - t_0) y \) for some \( t_0 \in [0, 1] \). Thus for any \( m', k > 0 \), there exist constants \( C, m > 0 \) such that when \( t > 1 \),
\[|F_t(x) - F_t(Y)| \overset{(a)}{=} \frac{2}{\sigma^2} \exp \left\{ - (x_0)^\top \frac{h(\theta)}{\sigma^2} \left( (I_d - e^{-2h(\theta)t})^{-1} \right) x_0 \right\} x_0^\top (X - Y) \]
\[\overset{(b)}{\leq} e^{-c|x_0|^2} |X_0| Ce^{-ct} (1 + |x|) \quad (A.18)\]
\[\overset{(c)}{\leq} C \frac{1 + |x|^m}{(1 + |x'|^m)(1 + t)^k}, \]
where in step (a),
\[X_0 = t_0 X + (1 - t_0) Y = x' - t_0 f(t, x, \theta) - (1 - t_0) h(\theta)^{-1} g(\theta), \quad (A.19)\]
for some \( t_0 \in [0, 1] \). Step (b) uses Equations (A.12) and (A.15) and step (c) is by substituting in \( x \) in Equation (A.1) to be the \( X_0 \) in Equation (A.19). Combining Equations (A.13), (A.14), (A.16), and (A.18), we have for \( t > 1 \),

\[
| p_t(x, x', \theta) - p_\infty(x', \theta) | \leq C \frac{1 + |x|^m}{(1 + |x'|^m')(1 + t)^k}.
\] (A.20)

The proof of Equation (26) for the case \( i = 1, 2 \) and Equation (27) is the same as the proof for \( | p_t(x, x', \theta) - p_\infty(x', \theta) | \) above (i.e., one uses the decomposition in Equations (A.13) and (A.1) with different choices of \( x \)). The only challenge is establishing a bound for \( \nabla \theta e^{-h(\theta)t} \).

Differentiating Equation (A.21) with respect to \( \theta_i \), \( i \in \{1, \ldots, d\} \) yields an ODE for \( \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \):

\[
\frac{d}{dt} e^{-h(\theta)t} = -h(\theta)e^{-h(\theta)t}
\] (A.21)

with initial value \( I_d \). Differentiating Equation (A.21) with respect to \( \theta_i \), \( i \in \{1, \ldots, d\} \) yields an ODE for \( \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \):

\[
\frac{d}{dt} \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} = -\frac{\partial h(\theta)}{\partial \theta_i} e^{-h(\theta)t} - h(\theta) \frac{\partial}{\partial \theta_i} e^{-h(\theta)t},
\] (A.22)

with initial value 0. Using an integrating factor yields

\[
\frac{d}{dt} \left( e^{h(\theta)t} \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right) = -e^{h(\theta)t} \frac{\partial h(\theta)}{\partial \theta_i} e^{-h(\theta)t},
\]

and thus

\[
\frac{\partial}{\partial \theta_i} e^{-h(\theta)t} = e^{-h(\theta)t} \int_0^t e^{h(\theta)s} \frac{\partial h(\theta)}{\partial \theta_i} e^{-h(\theta)s} ds.
\] (A.23)

Since \( e^{h(\theta)t} \) is invertible for any \( t \), we know that the matrices \( e^{h(\theta)s} \frac{\partial h(\theta)}{\partial \theta_i} e^{-h(\theta)s} \) and \( \frac{\partial h(\theta)}{\partial \theta_i} \) are similar and thus their eigenvalues are the same, which implies that their spectral norm are also the same.

We, therefore, can show that

\[
\left| \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right| \leq C \left| e^{-h(\theta)t} \right| \int_0^t \left| \frac{\partial h(\theta)}{\partial \theta_i} \right| ds \leq Ce^{-ct} t,
\] (A.24)

where step (a) is by the bound for \( \nabla \theta h(\theta) \) in Assumption 2.1. Using the same method, we also can show that

\[
\left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} e^{-h(\theta)t} \right| \leq C \left| e^{-h(\theta)t} \right| \int_0^t \left| \frac{\partial h(\theta)}{\partial \theta_i \partial \theta_j} \right| ds \leq Ce^{-ct} t, \ \forall i, j \in \{1, \ldots, d\}.
\] (A.25)

\[\text{Here we use the fact that } \frac{\partial}{\partial y} e^{Ay} = Ae^{Ay}.\]
Proof of (iv). The first part of Equation (28) follows from the fact that $X_t^\theta$ has a multivariate normal distribution whose mean and variance are uniformly bounded. Equation (28) is obvious when $t = 0$. For $t > 0$, as we know $\Sigma_t(\theta)$ is positive definite for $t > 0$, thus the random variable

$$Y := \Sigma_t^{-1/2}(\theta)(X_t^\theta - f(t, x, \theta))$$

(A.26)

has a $d$-dimensional standard normal distribution, where $\Sigma_t^{-1/2}(\theta)$ denotes the square root matrix of $\Sigma_t(\theta)$. Since for any $m > 0$, there exists a $C_m > 0$ such that $E|Y|^m = C_m < \infty$.

$$E_x|X_t^\theta|^m = E_x\left|\frac{1}{\Sigma_t}(\theta)Y + f(t, x, \theta)\right|^m \leq C\left(\left|\frac{1}{\Sigma_t}(\theta)E_x|Y|^m + |f(t, x, \theta)|^m\right|^a \leq C(1 + |x|^m),

(A.27)

where step (a) is by the uniform bound for $g(\theta)$ and $h(\theta)$ in Assumption 2.1. For the second part of Equation (28), we use Equation (A.27) to develop the following bound:

$$E_{x, \tilde{x}}|X_t^\theta|^m \leq 2|x|^m + 2E_{x, \tilde{x}}\left|\int_0^t \left|e^{-h(\theta)(t-s)}\left(V_\theta g(\theta) - V_\theta h(\theta)X_s^\theta\right)\right|ds\right|^m$$

(a)

$$\leq 2|x|^m + C_mE_{x, \tilde{x}}\left|\int_0^t e^{-c(t-s)}(1 + |X_s^\theta|)ds\right|^m$$

$$\leq 2|x|^m + C_mE_x\left|\int_0^t \frac{e^{cs}}{e^{ct} - 1}(1 + |X_s^\theta|)ds\right|^m \leq C_m(1 + |x|^m + |\tilde{x}|^m),$$

(A.28)

where step (a) is by Assumption 2.1 and the fact

$$\lambda_{\max}(e^{-h(\theta)(t-s)}) = e^{-\lambda_{\max}(h(\theta)(t-s))} \leq e^{-c(t-s)}$$

(A.29)

and step (b) is by Jensen’s inequality. In particular, let $p(s) = \frac{1}{c e^{cs}}$ and we have $\int_0^t p(s)ds = 1$, and therefore, $p(s)$ is a probability density function on $[0, t]$. By Jensen’s inequality,

$$\left|\int_0^t (1 + |X_s^\theta|)p(s)ds\right|^m \leq \int_0^t (1 + |X_s^\theta|)^m p(s)ds,$$

(A.30)

which we have used in step (b) of Equation (A.28).

Proof of (v). For Equation (29), the conclusion for $t = 0$ is trivial. When $t > 0$, by Equations (15) and (28), we have for any polynomial bounded function $f$ that

$$|E_x f(X_t^\theta)| \leq E_x |f(X_t^\theta)| \leq C E_x (1 + |X_t^\theta|^m) \leq C(1 + |x|^m).$$

(A.31)
For the derivatives, we will use the dominated convergence theorem. By Equation (A.9), we have

$$E_x f(X^\theta_t) = \int_{\mathbb{R}^d} f(f(t, x, \theta) + x') \frac{1}{\sqrt{(2\pi)^d |\Sigma_t(\theta)|}} \exp \left\{ -\frac{1}{2} x'^T \Sigma^{-1}_t(\theta) x' \right\} dx'. \quad (A.32)$$

Let $Z^\theta_t$ denote a normal distribution

$$Z^\theta_t \sim N(0, \Sigma_t(\theta))$$

and then

$$E_x f(X^\theta_t) = E f(f(t, x, \theta) + Z^\theta_t) = E f(e^{-h(\theta)t} x + h(\theta)^{-1} (I_d - e^{-h(\theta)t}) g(\theta) + Z^\theta_t). \quad (A.33)$$

For $\nabla_x E_x f(X^\theta_t)$, we change the order of $\nabla_x$ and $E_x$ and obtain for $t \in (0, 1]$

$$E \left| \nabla_x f(e^{-h(\theta)t} x + h(\theta)^{-1} (I_d - e^{-h(\theta)t}) g(\theta) + Z^\theta_t) \right| = E \left| e^{-ct} \nabla_x f(X^\theta_t) \right| \leq C(1 + |x|^m). \quad (A.34)$$

Therefore, by DCT we have that

$$\left| \nabla_x E_x f(X^\theta_t) \right| = \left| E \nabla_x f(f(t, x, \theta) + Z^\theta_t) \right| = \left| E \nabla_x f(X^\theta_t) \right| \leq C(1 + |x|^m). \quad (A.35)$$

Similarly for $\nabla^2_x E_x f(X^\theta_t)$, we have for $t \in (0, 1]$

$$\left| \nabla^2_x E_x f(X^\theta_t) \right| = \left| E e^{-h(\theta)t} \nabla^2 f \left( e^{-h(\theta)t} x + h(\theta)^{-1} (I_d - e^{-h(\theta)t}) g(\theta) + Z^\theta_t \right) \right|$$

$$\leq C(1 + |x|^m). \quad (A.36)$$

Finally, for $\frac{\partial}{\partial \theta_i} \nabla^2_x E_x f(X^\theta_t)$, by Equation (A.24), we have for $t \in (0, 1]$ that

$$\left| \frac{\partial}{\partial \theta_i} \nabla^2_x E_x f(X^\theta_t) \right| \leq 2 \left| \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right| \left| E_x \nabla^2 f(X^\theta_t) \right| e^{-h(\theta)t} + \left| e^{-h(\theta)t} \frac{\partial}{\partial \theta_i} E_x \nabla^2 f(X^\theta_t) e^{-h(\theta)t} \right|$$

$$\leq C(1 + |x|^m) + e^{-ct} \left| \frac{\partial}{\partial \theta_i} \nabla^2_x E_x f(X^\theta_t) \right|. \quad (A.37)$$

Thus, it remains to prove a bound for $\frac{\partial}{\partial \theta_i} E_x f_0(X^\theta_t)$, where $f_0$ is any polynomial bounded function such that

$$|f_0(x)| + |\nabla f_0(x)| \leq C(1 + |x|^m), \quad \forall x \in \mathbb{R}^d.$$
In order to establish this result, we need a bound for $\nabla_{\theta} \Sigma_t^{-1}(\theta)$ when $t \in [0, 1]$. For $t \in (0, 1]$, 

$$
\frac{\partial}{\partial \theta_i} \Sigma_t^{-1}(\theta) = 2\sigma^2 \frac{\partial}{\partial \theta_i} \left[ h(\theta)(I_d - e^{-2h(\theta)t})^{-1} \right]
$$

\[= (a) 2\sigma^2 \left( \frac{\partial}{\partial \theta_i} h(\theta) \right) (I_d - e^{-2h(\theta)t})^{-1} + 2\sigma^2 (I_d - e^{-2h(\theta)t})^{-1} h(\theta) \left( \frac{\partial}{\partial \theta_i} e^{-2h(\theta)t} \right) (I_d - e^{-2h(\theta)t})^{-1}
\]

\[= 2\sigma^2 (I_d - e^{-2h(\theta)t})^{-1} \left[ I_d - e^{-2h(\theta)t} - 2e^{-2h(\theta)t} h(\theta)t \right] \left( \frac{\partial}{\partial \theta_i} h(\theta) \right) (I_d - e^{-2h(\theta)t})^{-1},
\]

(A.38)

where in step (a), we change the order of $(I_d - e^{-2h(\theta)t})^{-1}$ and $h(\theta)$ since $h(\theta)e^{h(\theta)t} = e^{h(\theta)t} h(\theta).$ For $t \in [0, 1]$, 0 is the only singular point for $\nabla_{\theta} \Sigma_t^{-1}(\theta)$. Therefore, to prove the uniform bound, it suffices to prove the limit exists when $t \to 0^+$. As $t \to 0^+$,

$$
I_d - e^{-2h(\theta)t} - 2e^{-2h(\theta)t} h(\theta)t = I_d - 2h(\theta)t - (I_d - 2h(\theta)t + 2h^2(\theta)t^2 + o(t^2)) = -2h^2(\theta)t^2 + o(t^2).
\]

(A.39)

Therefore,

$$
\lim_{t \to 0^+} \frac{\partial}{\partial \theta_i} \Sigma_t^{-1}(\theta) = 2\sigma^2 \lim_{t \to 0^+} (2h(\theta)t + o(t))^{-1} \left( -2h^2(\theta)t^2 + o(t^2) \right) \left( \frac{\partial}{\partial \theta_i} h(\theta) \right)
\]

\[= 2\sigma^2 \lim_{t \to 0^+} (2h(\theta) + o(1))^{-1} \left( -2h^2(\theta) + o(1) \right) \left( \frac{\partial}{\partial \theta_i} h(\theta) \right)
\]

\[= -16h(\theta) \left( \frac{\partial}{\partial \theta_i} h(\theta) \right) h^{-1}(\theta),
\]

(A.40)

which together with the bound for $h(\theta)$ from Assumption 2.1 yields

$$\left| \nabla_{\theta} \Sigma_t^{-1}(\theta) \right| \leq C, \quad t \in [0, 1].$$

(A.41)

We will now analyze $\frac{\partial}{\partial \theta_i} \mathbf{E}_x f_0(X_t^\theta)$ for $t \in (0, 1]$ using formula (A.32) and changing the order of $\frac{\partial}{\partial \theta_i}$ and $\mathbf{E}_x$. 

$$
\int_{\mathbb{R}^d} \left| \frac{\partial}{\partial \theta_i} \right| f_0(f(t, x, \theta) + x') \frac{1}{\sqrt{(2\pi)^d | \Sigma_t(\theta) |}} \exp \left\{ -x'^T \Sigma_t^{-1}(\theta) x' \right\} dx' 
\]

\[\leq \int_{\mathbb{R}^d} \left( \frac{\partial}{\partial \theta_i} f(t, x, \theta) \right)^T \nabla f_0(f(t, x, \theta) + x') \frac{1}{\sqrt{(2\pi)^d | \Sigma_t(\theta) |}} \exp \left\{ -x'^T \Sigma_t^{-1}(\theta) x' \right\} dx' 
\]

\[+ \int_{\mathbb{R}^d} f_0(f(t, x, \theta) + x') \frac{\partial}{\partial \theta_i} \frac{1}{\sqrt{(2\pi)^d | \Sigma_t(\theta) |}} \exp \left\{ -x'^T \Sigma_t^{-1}(\theta) x' \right\} dx' 
\]

\[+ 2 \int_{\mathbb{R}^d} f_0(f(t, x, \theta) + x') \frac{1}{\sqrt{(2\pi)^d | \Sigma_t(\theta) |}} \exp \left\{ -x'^T \Sigma_t^{-1}(\theta) x' \right\} x'^T \frac{\partial}{\partial \theta_i} (\Sigma_t^{-1}(\theta)) x' \right\} dx' 
\]
\[ \begin{aligned}
&\leq C_E \left| \nabla f_0(f(t,x,\theta) + Z^\theta_t) \right| + C_E \left| f_0(f(t,x,\theta) + Z^\theta_t) \right| + C_E \left| f_0(f(t,x,\theta) + Z^\theta_t)(Z^\theta_t)^\top Z^\theta_t \right| \\
&\leq C_E \left| \nabla f_0(X^\theta_t) \right| + E_x \left| f_0(X^\theta_t) \right| + E_x \left| f_2^2(X^\theta_t) \right| + E \left| Z^\theta_t \right|^4
\end{aligned} \]

where step (a) is by Equation (A.41) and the uniform bounds for \( g(\theta), h(\theta) \) and step (b) is by Equation (A.31) and the polynomial bounds for \( f_2^2, \nabla f_0 \).

Then, by the dominated convergence theorem,

\[ \left| \nabla_\theta E_x f_0(X^\theta_t) \right| \leq C(1 + |x|^m), \quad t \in (0, 1]. \]  

Combining Equations (A.37) and (A.43), we obtain the bound for \( \nabla_\theta \nabla_\theta^2 E_x f(X^\theta_t) \). The bound \( \nabla_\theta \nabla_\theta^2 E_x f(X^\theta_t) \) can be obtained using similar calculations, which concludes the proof of the proposition. 

**APPENDIX B: POISSON PDES**

In this section, we give the detailed proof of the regularities for the solutions of Poisson PDEs. We first show the proof of Lemma 3.3.

**Proof of Lemma 3.3.** We begin by proving that the integral (31) is finite. We divide Equation (31) into two terms:

\[ v^1(x, \tilde{x}, \theta) = (E_{Y \sim \pi_\theta} f(Y) - \beta) \int_0^\infty \left( \nabla_\theta E_{Y \sim \pi_\theta} f(Y) - E_{x, \tilde{x}} \nabla f(X^\theta_t) \tilde{X}^\theta_t \right)^\top dt \]

\[ = (E_{Y \sim \pi_\theta} f(Y) - \beta) \left[ \int_0^\infty \left( \nabla_\theta E_{Y \sim \pi_\theta} f(Y) - \nabla_\theta E_{x, \tilde{x}} f(X^\theta_t) \right)^\top dt \right. \]

\[ + \left. \int_0^\infty \left( \nabla_\theta E_{x, \tilde{x}} f(X^\theta_t) - E_{x, \tilde{x}} \nabla f(X^\theta_t) \tilde{X}^\theta_t \right)^\top dt \right] \]

\[ =: v^{1,1}(x, \theta) + v^{1,2}(x, \tilde{x}, \theta). \]  

(B.1)

By Assumption 2.1 and Equation (25),

\[ \left| \int_{\mathbb{R}^d} f(x') \nabla^i_{\theta} p_{\infty}(x', \theta) dx' \right| \leq C \int \frac{1 + |x'|^m}{1 + |x'|^m'} dx' \leq C, \quad i = 0, 1, 2, \]  

(B.2)

where step (a) is by choosing \( m' > m + d \). Thus by dominated convergence theorem (DCT),

\[ \left| \nabla^i_{\theta} E_{Y \sim \pi_\theta} f(Y) \right| = \left| \int_{\mathbb{R}^d} f(x') \nabla^i_{\theta} p_{\infty}(x', \theta) dx' \right| \leq C, \quad i = 0, 1, 2. \]  

(B.3)
Similarly, we can bound $v^{1,1}$ as follows:
\[
|v^{1,1}(x, \theta)| \leq (a) C \int_0^1 \left( 1 + \left| \nabla \theta E_x f \left( X_t^\theta \right) \right| \right) dt + C \int_1^\infty \left( 1 + |x'|^m \right) \left| \nabla \theta p_\infty (x', \theta) - \nabla \theta p_1(x, x', \theta) \right| dx' dt \\
\leq (b) C + C \int_0^\infty \int_R \left( 1 + |x'|^m \right) \frac{1 + |x|^m'}{(1 + |x'|^m')(1 + t^2)} dx' dt \\
\leq (c) C(1 + |x|^m'),
\]
where steps (a) is by Assumption 2.1 and Equation (B.3), step (b) by Equations (26) and (29), and step (c) follows from selecting $m'' > m + d$. For $v^{1,2}$, by Assumption 2.1 and Equation (28), we have
\[
|E_{x,0} \nabla f \left( X_t^\theta \right) X_t^\theta| + E_x \left| \nabla f \left( X_t^\theta \right) \right|^2 \leq C < \infty. \tag{B.5}
\]
Thus by DCT, we have
\[
\nabla \theta E_x f \left( X_t^\theta \right) = E_{x,0} \nabla f \left( X_t^\theta \right) X_t^\theta ,
\tag{B.6}
\]
which together with Equation (23) derives
\[
\nabla \theta E_x f \left( X_t^\theta \right) - E_{x,\bar{x}} \nabla f \left( X_t^\theta \right) X_t^\theta = E_{x,0} \nabla f \left( X_t^\theta \right) X_t^\theta - E_{x,\bar{x}} \nabla f \left( X_t^\theta \right) X_t^\theta \\
= -E_{x} \nabla f \left( X_t^\theta \right) e^{-h(\theta)t} \bar{x} . \tag{B.7}
\]
Thus, $v^{1,2}$ satisfies the bound
\[
|v^{1,2}(x, \bar{x}, \theta)| \leq \left| (E_{Y \sim \pi_\theta f}(Y) - \beta) \int_0^\infty E_x \nabla f \left( X_t^\theta \right) e^{-h(\theta)t} \bar{x} dt \right| \\
\leq (a) C \int_0^\infty \left( 1 + E_x \left| X_t^\theta \right|^m \right) e^{-ct} dt \cdot |\bar{x}| \\
\leq (b) C \int_0^\infty \left( 1 + |x'|^{m''} \right) e^{-ct} dt \cdot |\bar{x}| \\
\leq C \left( 1 + |x'|^{m''} + |x'|^{m''} \right), \tag{B.8}
\]
where step (a) is by Assumption 2.1, Equation (B.3) and $\lambda_{max}(e^{-h(\theta)t}) \leq e^{-ct}$. Step (b) is by Equation (28).

Next we show that $v^1(x, \bar{x}, \theta)$ is differentiable with respect to $x, \bar{x}$, and $\theta$. We can prove this using a version of the dominated convergence theorem (see Theorem 2.27 in Folland (1999)), where it suffices to show that the derivative of the integrand is bounded by an integrable function. Using the same analysis as in Equation (B.8), we can show that
\[
\int_0^\infty e^{-h(\theta)t} E_x \nabla f \left( X_t^\theta \right) dt \leq C \int_0^\infty \left( 1 + E_x \left| X_t^\theta \right|^m \right) e^{-ct} dt \leq C \left( 1 + |x'|^{m''} \right). \tag{B.9}
\]
Therefore, by the dominated convergence theorem, we know that \( v^1 \) is differentiable with respect to \( \tilde{x} \). Furthermore, we can change the order of \( \nabla_{\tilde{x}} \) and the integral in \( v^1 \) and obtain

\[
|\nabla_{\tilde{x}} v^1(x, \tilde{x}, \theta)| = \left| (E_{Y \sim \pi_{\theta}} f(Y) - \beta) \int_0^\infty E_{x} \nabla f(X_t^\theta) e^{-h(\theta)t} dt \right| \leq C \left( 1 + |x|^m' \right),
\]

where step (a) is by Equations (B.3) and (B.9).

By Equations (26), (B.3), and the same approach as in Equation (B.4), we have

\[
|\nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y) \int_1^1 (\nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y) - \nabla_{\theta} E_{\tilde{x}} f(X_t^\theta)) dt + (E_{Y \sim \pi_{\theta}} f(Y) - \beta) \int_0^\infty (\nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y) - \nabla_{\theta} E_{\tilde{x}} f(X_t^\theta)) dt |
\leq C \int_1^1 \int_{R^d} |f(x')\nabla_{\theta} [p_{\infty}(x', \theta) - p_i(x, x', \theta)] dx' dt + C \int_1^1 \int_{R^d} |f(x')\nabla_{\theta} [p_{\infty}(x', \theta) - p_i(x, x', \theta)] dx' dt
\leq C (1 + |x|^m').
\]

By Equations (29) and (B.3),

\[
|\nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y) \int_0^1 \nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y) - \nabla_{\theta} E_{\tilde{x}} f(X_t^\theta) dt + (E_{Y \sim \pi_{\theta}} f(Y) - \beta) \int_0^1 \nabla_{\theta} E_{Y \sim \pi_{\theta}} f(Y) - \nabla_{\theta} E_{\tilde{x}} f(X_t^\theta) dt |
\leq C (1 + |x|^m').
\]

By Equations (B.11), (B.12), and DCT, we know that \( v^{1,1} \) is differentiable with respect to \( \theta \) and

\[
|\nabla_{\theta} v^{1,1}(x, \theta)| \leq C \left( 1 + |x|^m' \right).
\]

For \( \nabla_{\theta} v^{1,2} \), by Equation (B.7), we have for any \( i \in \{1, 2, \ldots, \ell \} \)

\[
\left| \int_0^\infty \frac{\partial}{\partial \theta_i} \left( \nabla_{\theta} E_{x} f(X_t^\theta) - E_{x, \tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta \right) dt \right|
= \left| \int_0^\infty E_{x} \nabla f(X_t^\theta) \left( \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right) \tilde{x} dt + \int_0^\infty \left( \frac{\partial}{\partial \theta_i} E_{x} \nabla f(X_t^\theta) \right) e^{-h(\theta)t} \tilde{x} dt \right|
\leq |\tilde{x}| \cdot \int_0^\infty \left| \frac{\partial}{\partial \theta_i} e^{-h(\theta)t} \right| \cdot \left| E_{x} \nabla f(X_t^\theta) \right| dt + |\tilde{x}| \cdot \int_0^\infty e^{-h(\theta)t} \cdot \left| E_{x} \frac{\partial}{\partial \theta_i} \nabla f(X_t^\theta) \right| dt
= : I_4 + I_5.
\]

where in step (a), we use

\[
\frac{\partial}{\partial \theta_i} E_{x} \nabla f(X_t^\theta) = E_{x} \frac{\partial}{\partial \theta_i} \nabla f(X_t^\theta),
\]

which is due to Equations (B.5) and (B.6).
By Equation (A.24),

\[ I_4 \leq C|\bar{x}| \left( \int_0^1 \left| \mathbf{E}_x \nabla f(X_1^\phi) e^{-ct} \right| dt + \int_1^{\infty} \left| \mathbf{E}_x \nabla f(X_1^\phi) e^{-ct} \right| dt \right) \]

\[ \leq C|\bar{x}| \left( 1 + \int_0^\infty e^{-ct} t \int_{\mathbb{R}^d} (1 + |x'|^m) |p_t(x, x', \theta) - p_\infty(x', \theta)| dx' dt \right. \]

\[ \left. +, \int_1^{\infty} e^{-ct} t \int_{\mathbb{R}^d} (1 + |x'|^m) |p_\infty(x', \theta)| dx' dt \right) \]

\[ \leq C(1 + |x|^m' + |\bar{x}|^m'), \tag{B.16} \]

where in step (a), we used Equation (29) and step (b) is by Equations (25), (26), and the same analysis as in Equations (B.3) and (B.4). Similarly,

\[ I_5 \leq C|\bar{x}| \left( \int_0^1 \left| \mathbf{E}_x \frac{\partial}{\partial \theta_i} \nabla f(X_1^\phi) \right| e^{-ct} dt + \int_1^{\infty} \left| \mathbf{E}_x \frac{\partial}{\partial \theta_i} \nabla f(X_1^\phi) \right| e^{-ct} dt \right) \]

\[ \leq C|\bar{x}| + C|\bar{x}| \cdot \int_0^\infty e^{-ct} \int_{\mathbb{R}^d} (1 + |x'|^m) \left| \frac{\partial}{\partial \theta_i} p_t(x, x', \theta) - \frac{\partial}{\partial \theta_i} p_\infty(x', \theta) \right| dx' dt \]

\[ + C|\bar{x}| \cdot \int_1^{\infty} e^{-ct} \int_{\mathbb{R}^d} (1 + |x'|^m) \left| \frac{\partial}{\partial \theta_i} p_\infty(x', \theta) \right| dx' dt \]

\[ \leq C(1 + |x|^m' + |\bar{x}|^m'). \tag{B.17} \]

Combining Equations (B.8), (B.14), (B.16), (B.17), and DCT, we know that \( u^{1,2} \) is differentiable with respect to \( \theta \) and for any \( i \in \{1, 2, \ldots, \ell\} \),

\[ \left| \frac{\partial u^{1,2}}{\partial \theta_i}(x, \bar{x}, \theta) \right| \leq \left| \frac{\partial}{\partial \theta_i} \mathbf{E}_{Y \sim \pi_\theta} f(Y) \int_0^{\infty} (e^{-h(\theta)t}) \mathbf{x} \nabla f(X_1^\phi) dt \right| \]

\[ + \left| \mathbf{E}_{Y \sim \pi_\theta} f(Y) - \beta \right| \cdot (I_4 + I_5) \leq C \left( 1 + |x|^m' + |\bar{x}|^m' \right), \tag{B.18} \]

which together with Equation (B.13) yields

\[ |\nabla_{\theta} u^1(x, \bar{x}, \theta)| \leq C \left( 1 + |x|^m' + |\bar{x}|^m' \right). \tag{B.19} \]

Similarly, by Equations (27)–(29),

\[ \left| \int_0^{\infty} \nabla_x \left( \nabla_{\theta} \mathbf{E}_{Y \sim \pi_\theta} f(Y) - \nabla_{\theta} \mathbf{E}_x f(X_1^\phi) \right) dt \right| \]

\[ = \left| \int_0^1 \nabla_x \nabla_{\theta} \mathbf{E}_x f(X_1^\phi) dx' dt \right| + \left| \int_1^{\infty} \int_{\mathbb{R}^d} f(x') \nabla_x \nabla_{\theta} p_t(x, x', \theta) dx' dt \right| \]

\[ \leq C \left( 1 + |x|^m' \right) \]
and

\[ \left| \int_0^\infty \nabla_x (\mathbf{E}_x f(X_t^\varphi) - \mathbf{E}_{x,\varphi} \nabla f(X_t^\varphi) X_t^\varphi) \, dt \right| \leq \left| \int_0^\infty \nabla_x (\mathbf{E}_x \nabla f(X_t^\varphi)) e^{-\kappa t} \, \xi \, dt \right| \leq C(1 + |x|^m + |\varphi|^m). \]

By DCT and Equation (B.3),

\[
\begin{align*}
|\nabla_x v^{1,1}(x, \theta)| &= \left| (\mathbf{E}_{Y \sim \pi_\theta} f(Y) - \beta) \int_0^\infty \int_{\mathbb{R}^d} f(x') \nabla_x \mathbf{v} \mathbf{p}_t(x, x', \theta) \, dx' \, dt \right| \leq C \left( 1 + |x|^m \right), \\
|\nabla_x v^{1,2}(x, \varphi, \theta)| &= \left| (\mathbf{E}_{Y \sim \pi_\theta} f(Y) - \beta) \int_0^\infty \nabla_x \mathbf{E}_x f(X_t^\varphi) e^{-\kappa(t)} \, \xi \, dt \right| \leq C \left( 1 + |x|^m + |\varphi|^m \right).
\end{align*}
\]

Then, for \( \nabla_x^2 v^1(x, \varphi, \theta) \), we have

\[
\left| \int_0^\infty \nabla_x^2 (\mathbf{E}_x f(Y) - \mathbf{E}_{x,\varphi} \nabla f(X_t^\varphi)) \, dt \right| \\
= \left| \int_0^1 \nabla_x^2 \mathbf{E}_x f(X_t^\varphi) \, dx' \, dt \right| + \left| \int_0^\infty \int_{\mathbb{R}^d} f(x') \nabla_x^2 \nabla_x \mathbf{v} \mathbf{p}_t(x, x', \theta) \, dx' \, dt \right| \\
\leq C \left( 1 + |x|^m \right),
\]

and

\[
\left| \int_0^\infty \nabla_x^2 (\mathbf{E}_x f(X_t^\varphi) - \mathbf{E}_{x,\varphi} \nabla f(X_t^\varphi) X_t^\varphi) \, dt \right| \\
= \left| \int_0^\infty \nabla_x^2 (\mathbf{E}_x \nabla f(X_t^\varphi)) e^{-\kappa t} \, \xi \, dt \right| \leq C(1 + |x|^m + |\varphi|^m).
\]

By DCT and Equation (B.3),

\[
\begin{align*}
|\nabla_x^2 v^{1,1}(x, \theta)| &= \left| (\mathbf{E}_{Y \sim \pi_\theta} f(Y) - \beta) \int_0^\infty \int_{\mathbb{R}^d} f(x') \nabla_x^2 \nabla_x \mathbf{v} \mathbf{p}_t(x, x', \theta) \, dx' \, dt \right| \leq C \left( 1 + |x|^m \right), \\
|\nabla_x^2 v^{1,2}(x, \varphi, \theta)| &= \left| (\mathbf{E}_{Y \sim \pi_\theta} f(Y) - \beta) \int_0^\infty \nabla_x^2 (\mathbf{E}_x \nabla f(X_t^\varphi)) e^{-\kappa(t)} \, \xi \, dt \right| \leq C \left( 1 + |x|^m + |\varphi|^m \right).
\end{align*}
\]

Finally, we verify that \( v^1 \) is a solution to the PDE (32). Note that

\[
\int_0^\infty \mathbf{E}_{x,\varphi} \mathbf{E}_{X_s^\varphi, X_t^\varphi} G^1(X_t^\varphi, X_s^\varphi, \theta) \, dt \overset{(a)}{=} \int_0^\infty \mathbf{E}_{x,\varphi} \mathbf{E}_{X_s^\varphi, X_t^\varphi} G^1(X_t^\varphi, X_s^\varphi, \theta) \, dt \overset{(b)}{=} \int_s^\infty \mathbf{E}_{x,\varphi} \mathbf{E}_{X_s^\varphi, X_t^\varphi} G^1(X_t^\varphi, X_s^\varphi, \theta) \, dt \overset{(c)}{<} \infty,
\]

where step (a) is by the Markov property of the process \((X_t^\varphi, \tilde{X}_t^\varphi)\), step (b) by change of variables, and step (c) is by the convergence of \( v^1 \). By Fubini’s theorem,

\[
\mathbf{E}_{x,\varphi} v^1(X_t^\varphi, X_s^\varphi, \theta) = \mathbf{E}_{x,\varphi} \int_0^\infty \mathbf{E}_{X_s^\varphi, X_t^\varphi} G^1(X_t^\varphi, X_s^\varphi, \theta) \, dt = \int_0^\infty \mathbf{E}_{x,\varphi} \mathbf{E}_{X_s^\varphi, X_t^\varphi} G^1(X_t^\varphi, X_s^\varphi, \theta) \, dt.
\]
Combining Equations (B.22) and (B.23), we have that

\[
\frac{1}{s} \left[ E_{x, \bar{x}} u^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) - u^1(x, \bar{x}, \theta) \right] = \frac{1}{s} \left[ - \int_0^\infty E_{x, \bar{x}} E_{x_\bar{t}, \bar{x}_\bar{t}}^\theta G^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) dt + \int_0^\infty E_{x, \bar{x}} G^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) dt \right]
\]

\[
= \frac{1}{s} \left[ - \int_0^\infty E_{x, \bar{x}} G^1(x_{\bar{t}+s}, \bar{x}_{\bar{t}+s}, \theta) dt + \int_0^\infty E_{x, \bar{x}} G^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) dt \right]
\]

\[
= \frac{1}{s} \left[ - \int_0^s E_{x, \bar{x}} G^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) dt + \int_0^\infty E_{x, \bar{x}} G^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) dt \right]
\]

\[
= \frac{1}{s} \int_0^s E_{x, \bar{x}} G^1(x_\bar{t}, \bar{x}_\bar{t}, \theta) dt
\]

(B.24)

Let \( s \to 0^+ \). By the definition of the infinitesimal generator and since \( u^1(x, \bar{x}, \theta) \) is twice differentiable with respect to \( x \) and once differentiable with respect to \( \bar{x} \), \( u^1(x, \bar{x}, \theta) \) is the classical solution of the Poisson PDE (32).

Now we show the proof of Lemma 3.6.

**Proof of Lemma 3.6.** The proof is exactly the same as in Lemma 3.3 except for the presence of the dimension \( \bar{x} \) and \( \mathcal{L}_{\bar{x}} \). We first show that the integral in Equation (57) converges. Note that

\[
v^2(x, \bar{x}, \bar{x}, \theta) = \int_0^\infty E_{x, \bar{x}, \bar{x}} \left[ (E_{Y \sim \pi_\theta} f(Y) - f(X_\bar{t}^\theta)) \cdot (\nabla f(X_\bar{t}^\theta) X_\bar{t}^\theta) \right] dt
\]

\[
= \int_0^\infty \left( E_{Y \sim \pi_\theta} f(Y) - E_{\bar{x}} f(X_\bar{t}^\theta) \right) \cdot E_{x, \bar{x}} \left( \nabla f(X_\bar{t}^\theta) X_\bar{t}^\theta \right) \right] dt
\]

(B.25)

where step (a) is by the independence of \( X_\bar{t}^\theta \) and \( (X_\bar{t}^\theta, \bar{x}_\bar{t}^\theta) \).

We now prove a uniform bound for \( E_{x, \bar{x}} \nabla f(X_\bar{t}^\theta) \bar{x}_\bar{t}^\theta \) and then by the ergodicity of \( X_\bar{t}^\theta \) in Lemma 3.1, we can show that the integrals converge.

\[
\left| E_{x, \bar{x}} \nabla f \left( X_\bar{t}^\theta \right) \bar{x}_\bar{t}^\theta \right| = \left| E_{x, \bar{x}} \nabla f \left( X_\bar{t}^\theta \right) \bar{x}_\bar{t}^\theta - \nabla_\theta f(x_\theta) + \nabla_\theta f(x_\theta) \right|
\]

\[
\leq (a) \left| \nabla_\theta f \left( X_\bar{t}^\theta \right) e^{-h(\theta)\bar{x}} \right| + \left| \nabla_\theta f(x_\theta) \right|
\]

(B.26)

where step (a) is by Equation (B.7). Therefore, for any \( t \in [0, 1] \), we can conclude

\[
\left| E_X \nabla f \left( X_\bar{t}^\theta \right) e^{-h(\theta)\bar{x}} \right| + \left| \nabla_\theta E_X f \left( X_\bar{t}^\theta \right) \right| \leq C \left( 1 + |x|^m + |ar{x}|^m \right),
\]

(B.27)
where we have used Assumption 2.1 and Equation (29). For \( t > 1 \), we have
\[
\left| E_\theta \nabla f (X_t^\theta) e^{-h^\theta(t)\tilde{x}} \right| + \left| \nabla_\theta E_\xi f (X_t^\xi) \right| \\
\stackrel{(a)}{\leq} C \left( 1 + E_\xi \left| X_t^\xi \right|^m \right) \cdot |\tilde{x}| + C \int_{\mathbb{R}^d} \left( 1 + |x'|^m \right) \left| \nabla_\theta p_t(x, x', \theta) - \nabla_\theta p_\infty(x', \theta) \right| dx' \\
+ C \int_{\mathbb{R}^d} \left( 1 + |x'|^m \right) \left| \nabla_\theta p_\infty(x', \theta) \right| dx' \\
\stackrel{(b)}{\leq} C \left( 1 + |x|^m' + |\tilde{x}|^m' \right),
\]
where step (a) uses Assumption 2.1 and step (b) uses Proposition 3.1 and the same calculations as in Equations (B.3) and (B.4). Combining Equations (B.27) and (B.28), we have for any \( t \geq 0 \),
\[
\left| E_{x, \tilde{x}} \nabla f (X_t^\tilde{x}) \overline{X}_t^\tilde{x} \right| \leq C \left( 1 + |x|^m' + |\tilde{x}|^m' \right).
\]
(B.29)

Thus, by Equation (B.29) and the same derivation as in Equation (B.4), we have
\[
|u^2(x, \tilde{x}, \tilde{x}, \theta)| \leq C \left( 1 + |x|^m' + |\tilde{x}|^m' \right) \cdot \int_0^\infty \left| E_{x, \tilde{x}} \nabla f (\overline{X}_t^\tilde{x}) - E_{Y \sim \pi_\theta} f(Y) \right| dt \\
\leq C \left( 1 + |x|^m' + |\tilde{x}|^m' + |\tilde{x}|^m' \right).
\]
(B.30)

We next show that \( u^2(x, \tilde{x}, \tilde{x}, \theta) \) is differentiable with respect to \( x, \tilde{x}, \tilde{x}, \theta \). Similar to Lemma 3.3, we first change the order of differentiation and integration and show the corresponding integral exists. Then, we apply DCT to prove that the differentiation and integration can be interchanged. For the ergodic process \( \overline{X}_t^\tilde{x}, \) by Equations (B.29), (27), and (29), we have the bounds
\[
\int_0^\infty \int_{\mathbb{R}^d} |f(x') \nabla_\tilde{x} p_t(\tilde{x}, x', \theta)| dx' \cdot \left| E_{x, \tilde{x}} \nabla f (\overline{X}_t^\tilde{x}) \overline{X}_t^\tilde{x} \right| dt \leq C \left( 1 + |x|^m' + |\tilde{x}|^m' + |\tilde{x}|^m' \right),
\]
\[
\int_0^\infty \int_{\mathbb{R}^d} \left| f (x') \nabla_\tilde{x}^2 p_t(\tilde{x}, x', \theta) \right| dx' \cdot \left| E_{x, \tilde{x}} \nabla f (\overline{X}_t^\tilde{x}) \overline{X}_t^\tilde{x} \right| dt \leq C \left( 1 + |x|^m' + |\tilde{x}|^m' + |\tilde{x}|^m' \right),
\]
(B.31)
and thus by the DCT,
\[
\sum_{i=1}^2 \left| \nabla_\tilde{x}^i u^2(x, \tilde{x}, \tilde{x}, \theta) \right| \leq C \left( 1 + |x|^m' + |\tilde{x}|^m' + |\tilde{x}|^m' \right).
\]
(B.32)
To address $V_x u^2, \nabla_x^2 u^2$, we first note that for any $i, j \in \{1, 2, \ldots, d\}$,

$$
\left| V_x E_{x,\xi} \nabla f (X_i) X_i \right| \leq \left| V_x E_x \nabla f (X_i) e^{-h(\theta)\xi} \right| + \left| \nabla_x V_\xi f (X_i) \right| \leq C \left( 1 + |x|^m + |\xi|^m \right),
$$

and Equation (B.38)

$$
\frac{\partial^2}{\partial x_i \partial x_j} E_{x,\xi} \nabla f (X_i) X_i \leq \left| \frac{\partial^2}{\partial x_i \partial x_j} E_x \nabla f (X_i) e^{-h(\theta)\xi} \right| + \left| \frac{\partial^2}{\partial x_i \partial x_j} V_\xi f (X_i) \right| \leq C \left( 1 + |x|^m + |\xi|^m \right),
$$

where in step (a), we use Equation (27) when $t > 1$, and Equation (29) for $t \in [0, 1]$. Thus, we have $\forall i, j \in \{1, 2, \ldots, d\}$

$$
\int_0^\infty \left| E_{Y - \pi_0} f(Y) - E_{x,\xi} f (X_i) \right| \cdot \left| V_x E_{x,\xi} \nabla f (X_i) X_i \right| dt \leq C \left( 1 + |x|^m + |\xi|^m + |\tilde{x}|^m \right),
$$

$$
\int_0^\infty \left| E_{Y - \pi_0} f(Y) - E_{x,\xi} f (X_i) \right| \cdot \left| \frac{\partial^2}{\partial x_i \partial x_j} E_{x,\xi} \nabla f (X_i) X_i \right| dt \leq C \left( 1 + |x|^m + |\xi|^m + |\tilde{x}|^m \right).
$$

Then by DCT,

$$
\sum_{i=1}^2 \left| V_i u^2(x, \tilde{x}, \tilde{\xi}, \theta) \right| \leq C \left( 1 + |x|^m + |\xi|^m + |\tilde{x}|^m \right).
$$

Then for $V_\theta u^2$, first we have for any $i \in \{1, 2, \ldots, \ell\}$

$$
\left| \frac{\partial}{\partial \theta_i} E_{x,\xi} \nabla f (X_i) X_i \right| \leq \left| \left( \frac{\partial}{\partial \theta_i} E_x \nabla f (X_i) \right) e^{-h(\theta)\xi} \right| + \left| E_x \nabla f (X_i) \left( \frac{\partial}{\partial \theta_i} e^{-h(\theta)\xi} \right) \right| + \left| \frac{\partial}{\partial \theta_i} V_\xi f (X_i) \right| \leq C \left( 1 + |x|^m + |\xi|^m \right),
$$

where in step (a), we use Equation (A.24) and the same analysis as in Equation (B.17). Thus,

$$
\left| \int_0^\infty \frac{\partial}{\partial \theta_i} \left( [E_{Y - \pi_0} f(Y) - E_{x,\xi} f (X_i)] \cdot E_{x,\xi} \nabla f(X_i) X_i \right) dt \right|
\leq \int_0^\infty \int_{\mathcal{R}^d} \left| f(x') \frac{\partial}{\partial \theta_i} (p_\infty(\tilde{x}', \theta) - p_\infty(\tilde{x}, \tilde{x}', \theta)) \right| d\tilde{x}' \cdot \left| E_{x,\xi} \nabla f (X_i) \right| dt
+ \int_0^\infty \int_{\mathcal{R}^d} \left| f(x') (p_\infty(\tilde{x}', \theta) - p_\infty(\tilde{x}, \tilde{x}', \theta)) \right| d\tilde{x}' \cdot \left| \frac{\partial}{\partial \theta_i} E_{x,\xi} \nabla f (X_i) \right| dt
\leq C \left( 1 + |x|^m + |\tilde{x}|^m + |\tilde{x}|^m \right),
$$

which together with the DCT derives

$$
\left| V_\theta u^2(x, \tilde{x}, \tilde{\xi}, \theta) \right| \leq C \left( 1 + |x|^m + |\tilde{x}|^m + |\tilde{x}|^m \right).
$$
Finally, note that

\[
\left| \int_0^\infty \nabla_x \left( \left[ \mathbb{E}_{Y \sim \pi_\theta} f(Y) - \mathbb{E}_{\tilde{x}} f(\tilde{X}_t^\theta) \right] \cdot \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta \right) dt \right|
\]
\[
\leq \int_0^\infty \left| \mathbb{E}_{Y \sim \pi_\theta} f(Y) - \mathbb{E}_{\tilde{x}} f(\tilde{X}_t^\theta) \right| \cdot \left| \nabla_x \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) \tilde{X}_t^\theta \right| dt
\]
\[
\leq C \int_0^\infty \left| \mathbb{E}_{Y \sim \pi_\theta} f(Y) - \mathbb{E}_{\tilde{x}} f(\tilde{X}_t^\theta) \right| \cdot \left| \mathbb{E}_{x,\tilde{x}} \nabla f(X_t^\theta) e^{-\gamma t} \right| dt
\]
\[
\leq C \left( 1 + |x|^{m'} + |\tilde{x}|^{m'} \right)
\]
\begin{equation}
(B.39)
\end{equation}

and then by DCT,

\[
|\nabla_x u^2(x, \tilde{x}, \bar{x}, \theta)| \leq C \left( 1 + |x|^{m'} + |\tilde{x}|^{m'} \right).
\begin{equation}
(B.40)
\end{equation}

By the same calculations as in Equation (B.24), we know that \(u^2\) is the classical solution of PDE (58) and the bound (60) holds. \(\square\)