Statistical Spatially Inhomogeneous Diffusion Inference

Yinuo Ren[1], **Yiping Lu**[2], **Lexing Ying**[1,3], **Grant M. Rotskoff**[1,4]
[1]Insitute for Computational and Mathematical Engineering (ICME), Stanford University
[2]Courant Institute of Mathematical Sciences, New York University
[3]Department of Mathematics, Stanford University
[4]Department of Chemistry, Stanford University
{yinuoren, lexing, rotskoff}@stanford.edu, yiping.lu@nyu.edu

Abstract

Inferring a diffusion equation from discretely-observed measurements is a statistical challenge of significant importance in a variety of fields, from single-molecule tracking in biophysical systems to modeling financial instruments. Assuming that the underlying dynamical process obeys a d-dimensional stochastic differential equation of the form

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{b}(\boldsymbol{x}_t)\mathrm{d}t + \Sigma(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{w}_t,$$

we propose neural network-based estimators of both the drift $\boldsymbol{b}$ and the spatially-inhomogeneous diffusion tensor $D = \Sigma\Sigma^T/2$ and provide statistical convergence guarantees when $\boldsymbol{b}$ and D are s-Hölder continuous. Notably, our bound aligns with the minimax optimal rate $N^{-\frac{2s}{2s+d}}$ for nonparametric function estimation even in the presence of correlation within observational data, which necessitates careful handling when establishing fast-rate generalization bounds. Our theoretical results are bolstered by numerical experiments demonstrating accurate inference of spatially-inhomogeneous diffusion tensors.

Introduction

The dynamical evolution of a wide variety of natural processes, from molecular motion within cells to atmospheric systems, involves an interplay between deterministic forces and noise from the surrounding environment. While it is possible to observe time series data from such systems, in general the underlying equation of motion is not known analytically. Stochastic differential equations offer a powerful and versatile framework for modeling these complex systems, but inferring the deterministic drift and diffusion tensor from time series data remains challenging, especially in high-dimensional settings. Among the many strategies proposed (Crommelin and Vanden-Eijnden 2011; Frishman and Ronceray 2020; Nickl 2022), there are few rigorous results on the optimality and convergence properties of estimators of, in particular, spatially-inhomogeneous diffusion tensors.

Many numerical algorithms have been proposed to infer the drift and diffusion, accommodating various settings, including one-dimensional (Sura and Barsugli 2002; Papaspiliopoulos et al. 2012; Davis and Buffett 2022) and multidimensional SDEs (Pokern, Stuart, and Vanden-Eijnden 2009; Frishman and Ronceray 2020; Crommelin and Vanden-Eijnden 2011). Also, the statistical convergence rate has been extensively studied for both the one-dimensional case (Dalalyan 2005; Dalalyan and Reiß 2006; Pokern, Stuart, and van Zanten 2013; Aeckerle-Willems and Strauch 2018) and the multidimensional cases (Van der Meulen, Van Der Vaart, and Van Zanten 2006; Dalalyan and Reiß 2007; van Waaij and van Zanten 2016; Nickl and Söhl 2017; Nickl and Ray 2020; Oga and Koike 2023; Nickl 2022). For parametric estimators using a Fourier or wavelet basis, the statistical limits of estimating the spatially-inhomogeneous diffusion tensor have been rigorously characterized (Hoffmann 1997, 1999a). However, strategies based on such decompositions do not scale to high-dimensional problems, which has motivated the investigation of neural networks as a more flexible representation of the SDE coefficients (Han, Jentzen, and E 2018; Rotskoff, Mitchell, and Vanden-Eijnden 2022; Khoo, Lu, and Ying 2021; Li et al. 2021).

Thus, we consider the nonparametric neural network estimator (Suzuki 2018; Oono and Suzuki 2019; Schmidt-Hieber 2020) as our ansatz function class, which has achieved great success in estimating SDE coefficients empirically (Xie et al. 2007; Zhang et al. 2018; Han, Jentzen, and E 2018; Wang et al. 2022; Lin, Li, and Ren 2023). We aim to build statistical guarantees for such neural network-based estimators. The most related concurrent work is (Gu et al. 2023), where the authors provide a convergence guarantee for the neural network estimation of the drift vector and the homogeneous diffusion tensor of an SDE by solving appropriate supervised learning tasks. However, their approach assumes that the data observed along the trajectory are independently and identically distributed from the stationary distribution. Additionally, the generalization bound used in (Gu et al. 2023) is not the fast rate generalization bound (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006), resulting in a suboptimal final guarantee. Therefore, we seek to bridge the gap between the i.i.d. setting and the non-i.i.d. ergodic setting using mixing conditions and extend the algorithm and analysis to the spatially-inhomogeneous diffusion estimation. We show that neural estimators have the ability to achieve standard minimax optimal nonparametric function estimation rates even when the data are non-i.i.d.

Contribution

In this paper, we construct a fast-rate error bound for estimating a multi-dimensional spatially-inhomogeneous diffusion process based on non-i.i.d ergodic data along a single trajectory. Our contributions are as follows:

- We derive for neural network-based diffusion estimators a convergence rate that matches the minimax optimal nonparametric function estimation rate for the s-Hölder continuous function class (Tsybakov and Zaiats 2009);
- Our analysis explores the β-mixing condition to address the correlation present among observed data along the trajectory, making our result readily applicable to a wide range of ergodic diffusion processes;
- We present numerical experiments, providing empirical support for our derived convergence rate and facilitating further applications of neural diffusion estimators in various contexts with theoretical assurance.

Our theoretical bound depicts the relationships between the error of nonparametric regression, numerical discretization, and ergodic approximation, and provides a general guideline for designing data-efficient, scale-minimal, and statistically-optimal neural estimators for diffusion inference.

Related Works

Inference of diffusion processes from data The problem of inferring the drift and diffusion coefficients of an SDE from data has been studied extensively in the literature. The setting with access to the whole continuous trajectory is studied by (Dalalyan and Reiß 2006, 2007; Strauch 2015, 2016; Nickl and Ray 2020; Rotskoff and Vanden-Eijnden 2019), in which the diffusion tensor can be exactly identified using quadratic variation arguments, and thus only the drift inference is considered. Many works focus on the numerical recovery of both the drift vector and the diffusion tensor in the more realistic setting when only discrete observations are available, including methods based on local linearization (Ozaki 1992; Shoji and Ozaki 1998), martingale estimating functions (Bibby and Sørensen 1995), maximum likelihood estimation (Pedersen 1995; Aït-Sahalia 2002), and Markov chain Monte Carlo (Elerian, Chib, and Shephard 2001). We refer readers to (Sørensen 2004; López-Pérez, Febrero-Bande, and González-Manteiga 2021) for an overview of parametric approaches. A spectral method that estimates the eigenpairs of the Markov semigroup operator is proposed in (Crommelin and Vanden-Eijnden 2011), and a nonparametric Bayesian inference scheme based on the finite element method is studied in (Papaspiliopoulos et al. 2012). As for the statistical convergence rate of the drift and diffusion inference, a line of pioneering works is by (Hoffmann 1997, 1999a,b), where the minimax convergence rate of the one-dimensional diffusion process is derived for Besov spaces and matched by adaptive wavelet estimators. Alternative analyses mainly follow a Bayesian methodology, with notable results by (Nickl and Söhl 2017; Nickl and Ray 2020; Nickl 2022) in both high- and low-frequency schemes.

Solving high-dimensional PDEs with deep neural networks The curse of dimensionality has stymied efforts to solve high-dimensional partial differential equations (PDEs) numerically. However, deep learning has demonstrated remarkable flexibility and adaptivity in approximating high-dimensional functions, which indeed has led to significant advances in computer vision and natural language processing. Recently, a series of works (Han, Jentzen, and E 2018; Yu et al. 2018; Karniadakis et al. 2021; Khoo, Lu, and Ying 2021; Long et al. 2018; Zang et al. 2020; Kovachki et al. 2021; Lu, Jin, and Karniadakis 2019; Li et al. 2022; Rotskoff, Mitchell, and Vanden-Eijnden 2022) have explored solving PDEs with deep neural networks, achieving impressive results for a diverse collection of tasks. These approaches rely on representing PDE solutions using neural networks, and various schemes propose different loss functions to obtain a solution. For instance, (Han, Jentzen, and E 2018) uses the Feynman-Kac formulation to convert PDE solving into a stochastic control problem, (Karniadakis et al. 2021) solves the PDE minimizing the strong form, while (Zang et al. 2020) solves weak formulations of PDEs with an adversarial approach.

Theoretical guarantees for neural network-based PDE solvers Statistical learning theory offers a powerful toolkit to prove theoretical convergence results for PDE solvers based on deep learning. For example, (Weinan and Wojtowytsch 2022; Chen, Lu, and Lu 2021; Marwah, Lipton, and Risteski 2021; Marwah et al. 2023) investigated the regularity of PDEs approximated by neural networks and (Nickl, van de Geer, and Wang 2020; Duan et al. 2021; Lu et al. 2021; Hütter and Rigollet 2021; Lu, Blanchet, and Ying 2022) consider the statistical convergence rate of various machine learning-based PDE solvers. However, most of these optimality results are based on concentration results that assume the sampled data are independent and identically distributed. This i.i.d. assumption is often violated in various financial and biophysical applications, for example, time series prediction, complex system analysis, and signal processing. Among many possible relaxations to this i.i.d. setting, the scenario, where data are drawn from a strong mixing process, has been widely adopted (Bradley 2005). Inspired by the first work of this kind (Yu 1994), many authors exploited a set of mixing concepts such as α-mixing (Zhang 2004; Steinwart and Christmann 2009), β- and ϕ-mixing (Mohri and Rostamizadeh 2008, 2010; Kuznetsov and Mohri 2017; Ziemann, Sandberg, and Matni 2022), and $\mathcal{C}$-mixing (Hang and Steinwart 2017). We refer readers to (Hang et al. 2016) for an overview of this line of research.

Notations We will use $\lesssim$ and $\gtrsim$ to denote the inequality up to a constant factor and $\asymp$ the equality up to a constant factor.

Definition 1 (Hölder space)**.** *We denote the Hölder space of order $s \in \mathbb{R}$ with constant $M > 0$ by $\mathcal{C}^s(\mathbb{R}^d, M)$,* i.e.

$$\mathcal{C}^s(\mathbb{R}^d, M) = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R} \Bigg| \sum_{|\boldsymbol{\alpha}|<s} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{|\boldsymbol{\alpha}|=\lfloor s \rfloor} \sup_{\boldsymbol{x}\neq\boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x}-\boldsymbol{y}|^{s-\lfloor s \rfloor}} < M \Bigg\}.$$

Problem Setting

Suppose we have access to a sequence of N discrete position snapshots $(\boldsymbol{x}_{k\tau})_{k=0}^{N}$ along a single trajectory $(\boldsymbol{x}_t)_{0\le t\le T}$, where the time step $\tau = T/N$ and $(\boldsymbol{x}_t)_{t\ge 0}$ is the solution to the following Itô stochastic differential equation:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{b}(\boldsymbol{x}_t)\mathrm{d}t + \Sigma(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{w}_t, \tag{1}$$

where $\boldsymbol{b} : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^d$, $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times r}$, and $(\boldsymbol{w}_t)_{t\ge 0}$ is an r-dimensional Wiener process. We refer the vector field $\boldsymbol{b}(\boldsymbol{x})$ as the drift vector, and define the diffusion tensor as $D(\cdot) = \frac{1}{2}\Sigma(\cdot)\Sigma(\cdot)^\top$. As noted in (Lau and Lubensky 2007), any interpolation between the Itô convention and other conventions for stochastic calculus can be transformed into the Itô convention by an additional term to the drift vector, and therefore, we work with the Itô convention throughout this paper[1].

Remark 1. *Our focus on the inhomogeneity in the space variable stems from the fact that when the SDE coefficients are time-dependent, it becomes very challenging to infer them from a singular observational trajectory,* i.e. *with only one observation at each time point and we would leave this case with multiple trajectories for future work.*

For simplicity, we will be working on $\Omega = [0,1)^d$ with periodic boundaries, *i.e.* the d-dimensional torus $\widetilde{\Omega} = \mathbb{R}^d/\mathbb{Z}^d$. Points on the torus $\widetilde{\Omega}$ are represented by $\widetilde{\boldsymbol{x}}$, where $\widetilde{\cdot}$ denotes the canonical map and $\boldsymbol{x} \in \mathbb{R}^d$ is a representative of the equivalence class $\widetilde{\boldsymbol{x}}$. The Borel σ-algebra on $\widetilde{\Omega}$ coincides with the sub-σ algebra of 1-periodic Borel sets of $\mathbb{R}^d$. We refer readers to (Papanicolau, Bensoussan, and Lions 1978) for further mathematical details of homogenization with tori. We further assume the drift and diffusion coefficients in (1) satisfy the following regularity assumptions:

Assumption 2a (Periodicity)**.** *$\boldsymbol{b}(\boldsymbol{x})$, $\Sigma(\boldsymbol{x})$, and $D(\boldsymbol{x})$ are 1-periodic for all variables.*

Remark 2. *This assumption is primarily for simplicity, and has been adopted in many previous works on the statistical inference of SDE coefficients,* e.g. *(Nickl and Ray 2020). This allows us to bypass the technicalities concerning boundary conditions, which might detract from our main contributions.*

Assumption 2b (Hölder-smoothness)**.** *Each entry $b_i(\boldsymbol{x}), \Sigma_{ij}(\boldsymbol{x}), D_{ij}(\boldsymbol{x}) \in \mathcal{C}^s(\mathbb{R}^d, M)$ for some $s \ge 2$ and $M > 0$.*

Assumption 2c (Uniform ellipticity)**.** *It holds that $r \ge d$ and there exists a constant $c > 0$ such that $D(\boldsymbol{x}) \succ cI$,* i.e. *$\sum_{i,j=1}^d D_{ij}(\boldsymbol{x})\xi_i\xi_j \ge c\|\xi\|^2$ for any $\xi \in \mathbb{R}^d$, holds uniformly for any $\boldsymbol{x} \in \mathbb{R}^d$.*

Remark 3. *This uniform ellipticity is commonly assumed across the analysis of the Fokker-Planck equation. It guarantees the Fokker-Planck equation has a unique strong solution with regularity properties that are essential for the analysis of asymptotic behavior and numerical approximation of the solution. We refer readers to (Stroock and Varadhan 1997; Bogachev et al. 2022) for more detailed discussions.*

Since $\boldsymbol{b}(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$ are 1-periodic, the process $(\boldsymbol{x}_t)_{t\ge 0}$ in (1) can thus be viewed as a process $(\widetilde{\boldsymbol{x}}_t)_{t\ge 0} := (\widetilde{\boldsymbol{x}_t})_{t\ge 0}$ on the torus $\widetilde{\Omega}$. Denote the transition kernel of the process $\boldsymbol{x}_t$ by $\mathbb{P}^t(\boldsymbol{x},\cdot) := \mathbb{P}(\boldsymbol{x}_t \in \cdot|\boldsymbol{x}_0 = \boldsymbol{x})$, the transition kernel of the corresponding process $\widetilde{\boldsymbol{x}}_t$ satisfies:

$$\widetilde{\mathbb{P}}^t(\widetilde{\boldsymbol{x}},\cdot) = \sum_{\boldsymbol{k}=(k_1,\cdots,k_d)\in\mathbb{Z}^d} \mathbb{P}^t\left(\boldsymbol{x}, \cdot + \sum_{i=1}^d k_i\boldsymbol{e}_i\right), \tag{2}$$

where $\boldsymbol{e}_i$ is the i-th standard basis vector in $\mathbb{R}^d$. When no confusion arises, we will use $\widetilde{\boldsymbol{x}}$ to denote its representative in the fundamental domain Ω in the following.

Spatially Inhomogeneous Diffusion Estimator

In this section, we aim to build neural estimators of both the drift and diffusion coefficients based on a sequence of N discrete observations $(\boldsymbol{x}_{k\tau})_{k=0}^{N}$ along a single trajectory of the SDE (1). A straightforward neural drift estimator allows us to subsequently construct a simple neural estimator of the diffusion tensor. In what follows, we introduce and prove the convergence of these neural estimators. Without loss of generality, we assume $\tau \le 1$ and $T \ge 1$, and denote the σ-algebra generated by all possible sequences $(\boldsymbol{x}_{k\tau})_{k=0}^{N}$ as $\mathcal{F}_{\tau,N}(\boldsymbol{b}, D)$.

Neural Estimators

We define $\mathcal{L}_T^{\boldsymbol{b}}(\hat{\boldsymbol{b}})$ and $\mathcal{L}_T^D(\hat{D})$ as the objective function for drift and diffusion estimation, respectively, by noticing that the ground truth drift vector $\boldsymbol{b}$ can be represented as the minimizer of the following objective function as the time step $\tau \to 0$ and the time horizon $T \to \infty$:

$$\mathcal{L}_T^{\boldsymbol{b}}(\hat{\boldsymbol{b}}; (\boldsymbol{x}_t)_{0\le t\le T}) := \frac{1}{T}\int_0^T \left\|\hat{\boldsymbol{b}}(\boldsymbol{x}_t) - \frac{1}{\tau}\Delta\boldsymbol{x}_t\right\|_2^2 \mathrm{d}t, \tag{3}$$

where $\Delta\boldsymbol{x}_t = \boldsymbol{x}_{t+\tau} - \boldsymbol{x}_t$. With the ground truth drift vector $\boldsymbol{b}$, the ground truth diffusion tensor can also be represented as the minimizer of the following objective function as $\tau \to 0$ and $T \to \infty$:

$$\mathcal{L}_T^D(\hat{D}; (\boldsymbol{x}_t)_{0\le t\le T}, \boldsymbol{b}) := \frac{1}{T}\int_0^T \left\|\hat{D}(\boldsymbol{x}_t) - \frac{(\Delta\boldsymbol{x}_t - \boldsymbol{b}(\boldsymbol{x}_t)\tau)(\Delta\boldsymbol{x}_t - \boldsymbol{b}(\boldsymbol{x}_t)\tau)^\top}{2\tau}\right\|_F^2 \mathrm{d}t, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

Based on the discussions in the last section, we will only estimate the value of $\hat{\boldsymbol{b}}$ and $\hat{D}$ in the fundamental domain Ω and then extend it to the whole space by periodicity. Therefore, using our data $(\boldsymbol{x}_{k\tau})_{k=0}^{N}$ as quadrature points, we approximate the objective function for drift estimation (3) as:

$$\tilde{\mathcal{L}}_N^{\boldsymbol{b}}(\hat{\boldsymbol{b}}; (\boldsymbol{x}_{k\tau})_{k=0}^{N}) := \frac{1}{N}\sum_{k=0}^{N-1}\left\|\frac{\boldsymbol{x}_{(k+1)\tau} - \boldsymbol{x}_{k\tau}}{\tau} - \hat{\boldsymbol{b}}(\widetilde{\boldsymbol{x}}_{k\tau})\right\|_2^2, \tag{5}$$

[1] The Itô convention along with others represent different methods to extend the Riemann integral to stochastic processes. Roughly speaking, Ito uses the left endpoint of the interval for functional value in the Riemann sum. We adopt the Itô convention due to several martingale properties it introduces which are mathematically convenient for statements and proofs.

Algorithm 1 Diffusion inference within function class $\mathfrak{G}$

1: Find the drift estimator

$$\hat{\boldsymbol{b}} := \operatorname*{arg\,min}_{\bar{\boldsymbol{b}} \in \mathfrak{G}^d} \tilde{\mathcal{L}}_N^{\boldsymbol{b}}(\bar{\boldsymbol{b}}; (\boldsymbol{x}_{k\tau})_{k=0}^N);$$

2: Find the diffusion estimator

$$\hat{D} := \operatorname*{arg\,min}_{\bar{D} \in \mathfrak{G}^{d\times d}} \tilde{\mathcal{L}}_N^{D}(\bar{D}; (\boldsymbol{x}_{k\tau})_{k=0}^N, \hat{\boldsymbol{b}}),$$

where $\hat{\boldsymbol{b}}$ is the drift estimator obtained in the first step as an approximation for the ground truth $\boldsymbol{b}$.

and the objective function for diffusion estimation (4) as

$$\tilde{\mathcal{L}}_N^D(\hat{D}; (\boldsymbol{x}_{k\tau})_{k=0}^N, \boldsymbol{b}) := \frac{1}{N}\sum_{k=0}^{N-1} \left\| \frac{(\Delta \boldsymbol{x}_{k\tau} - \boldsymbol{b}(\widetilde{\boldsymbol{x}}_{k\tau})\tau)(\Delta \boldsymbol{x}_{k\tau} - \boldsymbol{b}(\widetilde{\boldsymbol{x}}_{k\tau})\tau)^\top}{2\tau} - \hat{D}(\widetilde{\boldsymbol{x}}_{k\tau}) \right\|_F^2. \tag{6}$$

We will refer to $\tilde{\mathcal{L}}_N^{\boldsymbol{b}}(\hat{\boldsymbol{b}}; (\boldsymbol{x}_{k\tau})_{k=0}^N)$ and $\tilde{\mathcal{L}}_N^D(\hat{D}; (\boldsymbol{x}_{k\tau})_{k=0}^N, \boldsymbol{b})$ as the estimated empirical loss for drift and diffusion estimation, respectively.

We then parametrize the drift vector and the diffusion tensor within a *hypothesis function class* $\mathfrak{G}$ and solve for the estimators by optimizing the corresponding estimated empirical loss, as in Algorithm 1. Following foundational works including (Oono and Suzuki 2019; Schmidt-Hieber 2020; Chen et al. 2022), we adopt sparse neural networks $\mathfrak{N}(L, \boldsymbol{p}, S, M)$ as our hypothesis function class $\mathfrak{G}$, which is defined as follows. A neural network with depth L and width vector $\boldsymbol{p} = (p_0, \cdots, p_{L+1})$ has the following form $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ with

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}) = W_L(\sigma(W_{L-1}(\cdots \sigma(W_0 \boldsymbol{x} - \boldsymbol{w}_1)\cdots) - \boldsymbol{w}_L)), \tag{7}$$

where $W_i \in \mathbb{R}^{p_{i+1}\times p_i}$ are the weight matrices, $\boldsymbol{w}_i \in \mathbb{R}^{p_i}$ are the shift vectors, and $\sigma(\cdot)$ is the element-wise ReLU activation function. We also bound all parameters in the neural network by unity as in (Schmidt-Hieber 2020; Suzuki 2018).

Definition 3 (Sparse neural network)**.** *Let $\mathfrak{N}(L, \boldsymbol{p}, S, M)$ be the function class of ReLU-activated neural networks with depth L and width $\boldsymbol{p}$ that has at most S non-zero entries with the function value uniformly bounded by M and all parameters bounded by 1,* i.e.

$$\mathfrak{N}(L, \boldsymbol{p}, S, M) = \Bigg\{ f(\boldsymbol{x}) \text{ has the form of } (7) \Bigg| \sum_{i=0}^{L} \|W_i\|_0 + \sum_{i=1}^{L} \|\boldsymbol{w}_i\|_0 \le S, \|f\|_\infty \le M, \max_{i=0,\cdots,L} \|W_i\|_\infty \vee \max_{i=1,\cdots,L} \|\boldsymbol{w}_i\|_\infty \le 1 \Bigg\},$$

where $\|\cdot\|_0$ is the number of non-zero entries of a matrix (or a vector) and $\|\cdot\|_\infty$ is the maximum absolute value of a matrix (or a vector).

Since we are using the neural network for nonparametric estimation in $\Omega \subset \mathbb{R}^d$, we will assume $p_0 = d$ and $p_{L+1} = 1$ in the following discussion.

Ergodicity

Optimal convergence rates of neural network-based PDE solvers, as showcased in (Nickl, van de Geer, and Wang 2020; Lu et al. 2021; Gu et al. 2023), are typically established under the assumption of data independence. However, the presence of time correlations in the observational data $(\boldsymbol{x}_{k\tau})_{k=0}^N$ from a single trajectory significantly complicates the task of setting an upper bound for the convergence of the neural estimators obtained by Algorithm 1. In this context, we fully explore the ergodicity of the diffusion process, bound the ergodic approximation error by the β-mixing coefficient, and show that the exponential ergodicity condition, which is naturally satisfied by a wide range of diffusion processes, is sufficient for the fast rate convergence of the proposed neural estimators.

We first introduce the definition of exponential ergodicity:

Definition 4 (Exponential ergodicity (Down, Meyn, and Tweedie 1995))**.** *A diffusion process $(X_t)_{t\ge 0}$ with domain Ω is uniformly* exponential ergodic *if there exists a unique stationary distribution μ that for any $x \in \Omega$,*

$$\|\mathbb{P}^t(x, \cdot) - \mu\|_{\mathrm{TV}} \le M_\mu(x)\exp(-C_\mu t),$$

where $M_\mu(x), C_\mu > 0$.

As a direct consequence of (Papanicolau, Bensoussan, and Lions 1978, Theorem 3.2) and the compactness of the torus $\widetilde{\Omega}$, we have the following result:

Proposition 1 (Exponential ergodicity of $(\widetilde{\boldsymbol{x}}_t)_{t\ge 0}$)**.** *The diffusion process $(\widetilde{\boldsymbol{x}}_t)_{t\ge 0}$, the image of $(\boldsymbol{x}_t)_{t\ge 0}$ in (1) under the quotient map, is uniformly exponential ergodic with respect to a unique stationary distribution $\widetilde{\Pi}$ on the torus $\widetilde{\Omega}$ under Assumptions 2a, 2b, and 2c. Especially, there exist constants $M_{\widetilde{\Pi}}, C_{\widetilde{\Pi}} > 0$ that only depend on c, $\boldsymbol{b}$, and D, such that for any $\widetilde{\boldsymbol{x}} \in \Omega$,*

$$\|\widetilde{\mathbb{P}}^t(\widetilde{\boldsymbol{x}}, \cdot) - \widetilde{\Pi}\|_{\mathrm{TV}} \le M_{\widetilde{\Pi}}\exp(-C_{\widetilde{\Pi}} t).$$

See (Kulik 2017) for further discussions and required regularities for this property beyond the torus setting.

The ergodicity of stochastic processes is closely related to the notion of *mixing conditions*, which quantifies the "asymptotic independence" of random sequences. One of the most utilized mixing conditions for stochastic processes is the following β-mixing condition:

Definition 5 (β-mixing condition (Kuznetsov and Mohri 2017))**.** *The* β-mixing coefficient *of a stochastic process $(X_t)_{t\ge 0}$ with respect to a probability measure μ is defined as*

$$\beta(t; (X_t)_{t\ge 0}, \mu) := \sup_{s\ge 0} \mathbb{E}_{\mathcal{F}_0^t}\left[\|\mu - \mathbb{P}_{t+s}^\infty(\cdot|\mathcal{F}_0^t)\|_{\mathrm{TV}}\right],$$

where $\mathcal{F}_a^b$ is the σ-algebra generated by $(X_t)_{a\le t\le b}$, and $\mathbb{P}_a^b$ is the law of $(X_t)_{a\le t\le b}$. Especially, when $\beta(t; (X_t)_{t\ge 0}, \Pi) \le M_\beta \exp(-C_\beta t)$ for some constants $M_\beta, C_\beta > 0$, we say X_t is geometrically β-mixing *with respect to μ.*

By taking μ as the stationary distribution $\widetilde{\Pi}$ in the definition above, the proposition follows:

Proposition 2 (β-mixing condition of $(\widetilde{\boldsymbol{x}}_t)_{t\geq 0}$)**.** $\beta(t;(\widetilde{\boldsymbol{x}}_t)_{t\geq 0},\widetilde{\Pi}) \leq M_{\widetilde{\Pi}}\exp(-C_{\widetilde{\Pi}}t)$, i.e. $\widetilde{\boldsymbol{x}}_t$ *is geometrically* β*-mixing with respect to* $\widetilde{\Pi}$.

We will denote the pushforward of the invariant measure $\widetilde{\Pi}$ under the following inverse of the canonical map $\iota^{-1}:\widetilde{\Omega}\to\Omega$ also as $\widetilde{\Pi}$.

Convergence Guarantee

In this section, we describe the main upper bound for the neural estimators in Algorithm 1. We also present a theoretical guarantee for drift and diffusion estimation in Theorem 3 and 4, respectively. Our main result shows that estimating the drift and diffusion tensor can achieve the standard minimax optimal nonparametric function estimation convergence rate, even with non-i.i.d. data.

Due to the ergodic theorem (Kulik 2017, Theorem 5.3.3) under the exponential ergodicity condition and the property of Itô process, the *bias* part of the objective functions $\mathcal{L}_T^{\boldsymbol{b}}(\hat{\boldsymbol{b}};(\boldsymbol{x}_t)_{0\leq t\leq T})$ and $\mathcal{L}_T^D(\hat{D};(\boldsymbol{x}_t)_{0\leq t\leq T},\boldsymbol{b})$ for drift and diffusion estimation as defined in (3) and (4) converge to

$$\begin{aligned}\mathcal{L}_{\widetilde{\Pi}}^{\boldsymbol{b}}(\hat{\boldsymbol{b}}) &:= \mathbb{E}_{\widetilde{\boldsymbol{x}}\sim\widetilde{\Pi}}\left[\|\hat{\boldsymbol{b}}(\widetilde{\boldsymbol{x}})-\boldsymbol{b}(\widetilde{\boldsymbol{x}})\|_2^2\right]\\ \text{and}\quad \mathcal{L}_{\widetilde{\Pi}}^{D}(\hat{D}) &:= \mathbb{E}_{\widetilde{\boldsymbol{x}}\sim\widetilde{\Pi}}\left[\|\hat{D}(\widetilde{\boldsymbol{x}})-D(\widetilde{\boldsymbol{x}})\|_F^2\right],\end{aligned} \tag{8}$$

as $\tau\to 0$ and $T\to\infty$, which we will refer to as the *population loss* for drift and diffusion estimation, respectively. Our convergence guarantee is thus built on these population losses.

Theorem 3 (Upper bound for drift estimation in $\mathfrak{N}(L,\boldsymbol{p},S,M)$)**.** *Suppose the drift vector* $\boldsymbol{b}\in\mathcal{C}^s(\Omega,M)$*, and the hypothesis class* $\mathfrak{G}=\mathfrak{N}(L,\boldsymbol{p},S,M)$ *with*

$$K\asymp T^{\frac{d}{2s+d}},\quad L\lesssim\log K,\quad \|\boldsymbol{p}\|_\infty\lesssim K,\quad S\lesssim K\log K.$$

Then with high probability the minimizer $\hat{\boldsymbol{b}}$ *obtained by Algorithm 1 satisfies*

$$\mathbb{E}\left[\mathcal{L}_{\widetilde{\Pi}}^{b}(\hat{\boldsymbol{b}})\right]\lesssim T^{-\frac{2s}{2s+d}}\log^3 T+\tau,$$

where the expectation is taken over $(\boldsymbol{x}_{k\tau})_{k=0}^N\sim\mathcal{F}_{\tau,N}(\boldsymbol{b},D)$.

Theorem 4 (Upper bound for diffusion estimation in $\mathfrak{N}(L,\boldsymbol{p},S,M)$)**.** *Suppose the diffusion tensor* $D\in\mathcal{C}^s(\Omega,M)$*, and the hypothesis class* $\mathfrak{G}=\mathfrak{N}(L,\boldsymbol{p},S,M)$ *with*

$$K\asymp N^{\frac{d}{2s+d}},\quad L\lesssim\log K,\quad \|\boldsymbol{p}\|_\infty\lesssim K,\quad S\lesssim K\log K.$$

Then with high probability the minimizer $\hat{D}$ *obtained by Algorithm 1 satisfies*

$$\mathbb{E}\left[\mathcal{L}_{\widetilde{\Pi}}^{D}(\hat{D})\right]\lesssim N^{-\frac{2s}{2s+d}}\log^3 N+\tau+\frac{\log^2 N}{T}, \tag{9}$$

where the expectation is taken over $(\boldsymbol{x}_{k\tau})_{k=0}^N\sim\mathcal{F}_{\tau,N}(\boldsymbol{b},D)$.

Remark 4. *In this remark, we explain the meaning of each term in the convergence rate* (9)*:*

- *The term* $N^{-\frac{2s}{2s+d}}\log^3 N$ *matches the standard minimax optimal rate* $N^{-\frac{2s}{2s+d}}$ *up to an extra* $\mathrm{poly}(\log n)$ *factor. This is characteristic of performing nonparametric regression for* s*-Hölder continuous functions with* N *noisy observations (Tsybakov and Zaiats 2009). This is different from the drift estimation (Theorem 3) in which the nonparametric dependency is on* T *instead of* N *with a further* $\frac{\log^2 N}{T}$ *term which is discussed below.*
- *The term* τ *represents a bias term that arises due to the finite resolution of the observations* $(\boldsymbol{x}_{k\tau})_{t=0}^N$*. Specifically, this term encapsulates the error incurred while approximating the objective function* $\mathcal{L}_T^D(\hat{D};(\boldsymbol{x}_t)_{0\leq t\leq T},\boldsymbol{b})$ *by the estimated empirical loss* $\tilde{\mathcal{L}}_N^D(\hat{D};(\boldsymbol{x}_{k\tau})_{k=0}^N,\hat{\boldsymbol{b}})$ *with numerical quadrature and finite difference computations;*
- *The term* $\frac{\log^2 N}{T}$ *quantifies the error in approximating the population loss* $\mathcal{L}_{\widetilde{\Pi}}^D(\hat{D})$ *by the objective function* $\mathcal{L}_T^D(\hat{D};(\boldsymbol{x}_t)_{0\leq t\leq T},\boldsymbol{b})$ *by applying the ergodic theorem up to time horizon* T*. This term essentially signifies the portion of the domain that the trajectory has not yet traversed. Refs. (Hoffmann 1997, 1999a) only provide guarantee for* $\mathcal{L}_T^D(\hat{D};(\boldsymbol{x}_t)_{0\leq t\leq T},\boldsymbol{b})$ *and thus this term is not included.*

Proof Sketch

In this section, we omit the dependency of the losses on the data $(\boldsymbol{x}_{k\tau})_{k=0}^N$ for notational simplicity and unless otherwise stated, the expectation is taken over $(\boldsymbol{x}_{k\tau})_{k=0}^N\sim\mathcal{F}_{\tau,N}(\boldsymbol{b},D)$.

To obtain a unified proving approach for both drift and diffusion estimation, it is useful to think of our neural estimator as a function regressor with imperfect supervision signals. We consider an estimator $\hat{g}\in\mathfrak{G}$ of an arbitrary function g^0 as the ground truth obtained by minimizing over the estimated empirical loss

$$\tilde{\mathcal{L}}_N^{g^0}(\hat{g})=\frac{1}{N}\sum_{k=0}^{N-1}\left(g^0(\widetilde{\boldsymbol{x}}_{k\tau})+\Delta Z_{k\tau}-\hat{g}(\widetilde{\boldsymbol{x}}_{k\tau})\right)^2, \tag{10}$$

where the supervision signal is polluted by the noise given by $\Delta Z_{k\tau}=Z_{(k+1)\tau}-Z_{k\tau}$, with Z_t being an $\mathcal{F}_t$-adapted continuous semimartingale. Following Doob's decomposition, we write $Z_t=A_t+M_t$, where $(A_t)_{t\geq 0}$ is a continuous process with finite variation and is deterministic on $[k\tau,(k+1)\tau]$ conditioned on $\mathcal{F}_0^{k\tau}$ as

$$A_t=\sum_{k=0}^{N-1}\left(\mathbb{E}\left[Z_{t\wedge(k+1)\tau}\mid\mathcal{F}_{t\wedge k\tau}\right]-Z_{t\wedge k\tau}\right)$$

and $(M_t)_{t\geq 0}$ forms a local martingale as

$$M_t=\sum_{k=0}^{N-1}\left(Z_{t\wedge(k+1)\tau}-\mathbb{E}\left[Z_{t\wedge(k+1)\tau}\mid\mathcal{F}_{t\wedge k\tau}\right]\right).$$

The population loss $\mathcal{L}_{\widetilde{\Pi}}^{g^0}(\hat{g})$ can also be similarly defined as in (8). Additionally, we define the *empirical loss* for the estimator $\hat{g}$ as

$$\hat{\mathcal{L}}_N^{g^0}(\hat{g}) := \frac{1}{N}\sum_{k=0}^{N-1}\left(g^0(\widetilde{\boldsymbol{x}}_{k\tau}) - \hat{g}(\widetilde{\boldsymbol{x}}_{k\tau})\right)^2.$$

In our proof, we first show that as long as the following two conditions hold for the noise $(\Delta Z_{k\tau})_{k=0}^N$, the minimax optimal nonparametric function estimation rate would be achieved:

Assumption 6. *For any k, the continuous finite variation process $(A_t)_{t\geq 0}$ satisfies*

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1}(\Delta A_{k\tau})^2\right] \leq C_A\tau.$$

Assumption 7. *For some $\gamma \leq 1$, the local martingale $(M_t)_{t\geq 0}$ satisfies*

$$\max_k \left|\mathbb{E}\left[\Delta\langle M\rangle_{k\tau}\left|\mathcal{F}_0^{k\tau}\right.\right]\right| \leq C_M\tau^{-\gamma},$$

where $\langle\cdot\rangle$ denotes the quadratic variation.

Remark 5. *Based on the noise decomposition $\Delta Z_{k\tau} = \Delta A_{k\tau} + \Delta M_{k\tau}$, the term $\Delta A_{k\tau}$ can be intuitively understood as the bias of the data. This bias is caused by the numerical scheme employed for computing $g^0_{k\tau}$. On the other hand, the term $\Delta M_{k\tau}$ represents the martingale noise added to the data, which can be considered analogous to the i.i.d. noise in the common nonparametric estimation settings. Assumption 6 essentially implies that the estimator $\hat{g}$ is consistent, for its expectation converges to g^0 as $\tau \to 0$. Meanwhile, Assumption 7 assumes that the variance of the noise present in the data is at most of order $\mathcal{O}(\tau^{-1})$.*

To overcome the correlation of the observed data, we adopt the following *sub-sampling* technique as in (Yu 1994; Mohri and Rostamizadeh 2008; Hang and Steinwart 2017): For a sufficiently large $l \geq 1$ such that $N = nl$[2], we split the original N correlated samples $S^N := (\boldsymbol{x}_{k\tau})_{k=0}^N$ into l sub-sequences $S^n_{(a)} := (\boldsymbol{x}_{(kl+a)\tau})_{k=0}^{n-1}$ for $a = 0, \cdots, l-1$. The main idea of this technique is that under fast β-mixing conditions, each sub-sequence can be treated approximately as n i.i.d. samples from the distribution $\widetilde{\Pi}$ to which the classical generalization results may apply, with an error that can be controlled by the mixing coefficient via the following lemma:

Lemma 5 ((Kuznetsov and Mohri 2017, Proposition 2))**.** *Let h be any function on $\widetilde{\Omega}^n$ with $-M_1 \leq h \leq M_2$ for $M_1, M_2 \geq 0$. Then for any $0 \leq a \leq l-1$, we have*

$$\left|\mathbb{E}_{\widetilde{S}^n_{\widetilde{\Pi}}\sim\widetilde{\Pi}^{\otimes n}}\left[h(\widetilde{S}^n_{\widetilde{\Pi}})\right] - \mathbb{E}\left[h(\widetilde{S^n_{(a)}})\right]\right| \leq (M_0+M_1)n\beta(l\tau),$$

where the second expectation is taken over the sub-σ-algebra of $\mathcal{F}_0^T$ generated by the sub-sequence $S^n_{(a)} = (\boldsymbol{x}_{(kl+a)\tau})_{k=0}^{n-1}$ and $\widetilde{S^n_{(a)}} := (\widetilde{\boldsymbol{x}}_{(kl+a)\tau})_{k=0}^{n-1}$.

[2]Here we assume N is divisible by l without loss of generality.

Based on Lemma 5, we derive the following fast rate generalization bound via local Rademacher complexity arguments (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006). The proof is shown in Appendix.

Theorem 6. *Let $N = nl$. Suppose the localized Rademacher complexity satisfies*

$$\mathfrak{R}_N(\{\ell\circ g | g\in\mathfrak{G}, \mathbb{E}_\Pi[\ell\circ g] \leq r\}) \leq \phi(r),$$

where $\phi(r)$ is a sub-root function[3] *and $\mathfrak{R}_N(\mathfrak{F})$ is the Rademacher complexity of a function class $\mathfrak{F}$ with respect to N i.i.d. samples from the stationary distribution $\widetilde{\Pi}$,* i.e.

$$\mathfrak{R}_N(\mathfrak{F}) = \mathbb{E}\left[\sup_{f\in\mathfrak{F}}\frac{1}{N}\sum_{i=1}^N \sigma_i f(\widetilde{X}_i)\right], \tag{11}$$

where $(\widetilde{X}_i)_{i=1}^N \sim \widetilde{\Pi}^{\otimes N}, \boldsymbol{\sigma} \sim \mathrm{Unif}(\{\pm 1\}^N)$.

Let r^* be the unique solution to the fixed-point equation $\phi(r) = r$. Then for any $\delta > N\beta(l\tau)$ and $\epsilon > 0$, we have with probability $1-\delta$ for any $g \in \mathfrak{G}$,

$$\mathcal{L}_{\widetilde{\Pi}}^{g^0}(\hat{g}) \leq \frac{1}{1-\epsilon}\hat{\mathcal{L}}_N^{g^0}(\hat{g}) + \frac{176}{M^2\epsilon}r^* + \frac{(44\epsilon+104)M^2\log\left(\frac{l}{\delta'}\right)}{\epsilon n}, \tag{12}$$

where $\delta' = \delta - N\beta(l\tau)$.

Bias and noise in the objective function certainly affect the optimization. Thus, we need to seek an oracle-type inequality for the expectation of the population loss $\hat{\mathcal{L}}_N^{g^0}(\hat{g})$ over the data, which is proved in Appendix. The main technique is a uniform martingale concentration inequality (*cf.* Lemma 12).

Theorem 7. *Suppose $\mathfrak{G}$ is separable with respect to the L^∞ norm with ρ-covering number $\mathcal{N}(\rho, \mathfrak{G}, \|\cdot\|_\infty) \geq 2$. Then under Assumption 6 and 7, we have*

$$\mathbb{E}\left[\hat{\mathcal{L}}_N^{g^0}(\hat{g})\right] \leq \frac{1+\epsilon}{1-\epsilon}\inf_{\bar{g}\in\mathfrak{G}}\mathbb{E}\left[\hat{\mathcal{L}}_N^{g^0}(\bar{g})\right] + \frac{3C_A}{\epsilon}\tau + \frac{12C_M\log\mathcal{N}(\rho,\mathfrak{G},\|\cdot\|_\infty)}{\epsilon N\tau^\gamma} + 2\sqrt{\frac{4C_M\rho^2\log 2}{N\tau^\gamma}}.$$

Especially, when we choose the hypothesis class $\mathfrak{G}$ as the sparse neural network class $\mathfrak{N}(L, \boldsymbol{p}, S, M)$ and combine Theorem 6 and Theorem 7, we have the following theorem with the proof given in Appendix:

Theorem 8. *Suppose Assumption 6 and 7 are satisfied and the ground truth $g^0 \in \mathcal{C}^s(\Omega, M)$, and the hypothesis class $\mathfrak{G} = \mathfrak{N}(L, \boldsymbol{p}, S, M)$ with*

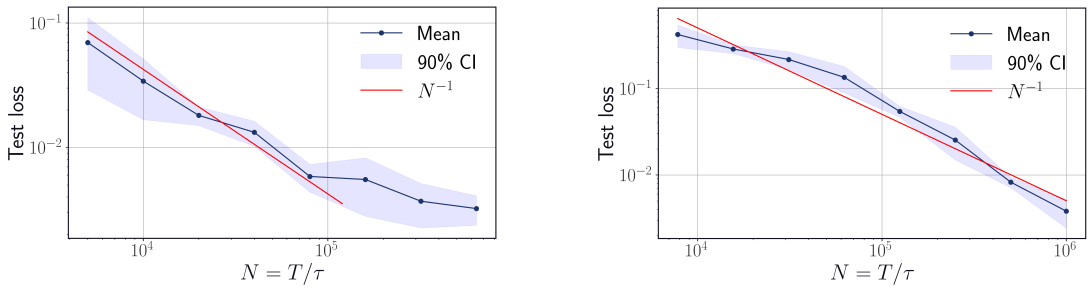
$$K \asymp \tilde{N}^{\frac{d}{2s+d}}, \quad L \lesssim \log K,$$
$$\|\boldsymbol{p}\|_\infty \lesssim K, \quad S \lesssim K\log K,$$

where $\tilde{N} = N(\tau^\gamma \wedge 1)$. Then with high probability the minimizer $\hat{g}$ obtained by minimizing the estimated empirical loss $\hat{\mathcal{L}}_N^{g^0}(\hat{g}; (\boldsymbol{x}_{k\tau})_{k=0}^N)$ satisfies

$$\mathbb{E}\left[\mathcal{L}_{\widetilde{\Pi}}^{g^0}(\hat{g})\right] \lesssim \tilde{N}^{-\frac{2s}{2s+d}}\log^3\tilde{N} + \tau + \frac{\log^2 N}{N\tau}. \tag{13}$$

With Theorem 8, the detailed proofs of Theorem 3 and 4 are given in Appendix.

[3]A sub-root function $\phi(r)$ is a function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ that is non-negative, non-decreasing function, satisfying that $\phi(r)/\sqrt{r}$ is non-increasing for $r > 0$.



(a) $\tau = 10^{-3}$ with varying N and T

(b) $T = 10^3$ with varying N and τ

Figure 1: Numerical results of the neural diffusion estimator are consistent with the scaling expected from the theoretical bound. We probe this by varying N and $T = N\tau$ with fixed lag time τ and also by varying N and $\tau = T/N$ with fixed time horizon T.

Experiments

In this section, we present numerical results on a two-dimensional example, to illustrate the accordance between our theoretical convergence rates and those of our proposed neural diffusion estimator. Consider the following SDE in $\mathbb{R}^2$:

$$\mathrm{d}\boldsymbol{x}_t = f(\boldsymbol{x}_t)\nabla f(\boldsymbol{x}_t)\mathrm{d}t + f(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{w}_t, \tag{14}$$

where

$$f(\boldsymbol{x}) = 1 + \frac{1}{2}\cos(2\pi(x_1 + x_2)),$$

i.e. $\boldsymbol{b}(\boldsymbol{x}) = f(\boldsymbol{x})\nabla f(\boldsymbol{x})$ and $D(\boldsymbol{x}) = \frac{f(\boldsymbol{x})^2}{2}I$, where I is the 2×2 identity matrix. Then it is straightforward to verify that this diffusion process satisfies Assumption 2a, 2b, and 2c with smoothness $s = \infty$. Our goal is to estimate the value of the function $f(\boldsymbol{x})$ within $\Omega = [0, 1)^2$. We employ Algorithm 1 for estimating both $\boldsymbol{b}(\boldsymbol{x})$ and $D(\boldsymbol{x})$ with separate neural networks and treat them entirely independently in the inference task. One may also prove that the stationary distribution $\widetilde{\Pi}$ of this diffusion process is given by the Lebesgue measure on the two-dimensional torus, which makes evaluating errors easier and more precise.

To impose the periodic boundary, we introduce an explicit regularization term to our training loss

$$\mathcal{L}_{\mathrm{per}}(\hat{g}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathrm{Unif}(\partial\Omega^2),\widetilde{\boldsymbol{x}}=\widetilde{\boldsymbol{y}}}\left[\left(\hat{g}(\boldsymbol{x}) - \hat{g}(\boldsymbol{y})\right)^2\right],$$

approximated by $\hat{\mathcal{L}}_{\mathrm{per}}(\hat{g})$ with 1000 pairs of random samples empirically. The final training loss is thus $\tilde{\mathcal{L}}^N(\hat{g}) + \lambda\hat{\mathcal{L}}_{\mathrm{per}}(\hat{g})$, where λ is a hyperparameter and $\hat{g}$ can be either $\hat{\boldsymbol{b}}$ or $\hat{D}$.

We first generate data using the Euler-Maruyama method with a time step $\tau_0 = 2 \times 10^{-5}$ up to $T_0 = 10^4$, and then sub-sample data at varying time steps τ and time horizons T for each experiment instance from this common trajectory. We use a ResNet as our neural network structure with two residual blocks, each containing a fully-connected layer with a hidden dimension of 1000. Test data are generated by randomly selecting 5×10^4 samples from another sufficiently long trajectory, which are shared by all experiment instances. The training process is executed on one Tesla V100 GPU.

According to our theoretical result (Theorem 4), the convergence rate of this implementation should be approximately of order $N^{-1} + \tau + T^{-1}$ up to $\log$ terms. We thus consider two schemes in our experiment. The first involves a fixed time step $\tau = 10^{-3}$ with an expected rate of $\tau + N^{-1}$, and the other maintains a fixed $T = 10^3$ with an expected rate of $N^{-1} + T^{-1}$. Each of the aforementioned instances is carried out five times. Figure 1 presents the mean values along with their corresponding confidence intervals from these runs. Additionally, reference lines indicating the expected convergence rate N^{-1} are shown in red. Both schemes roughly exhibit the exponential decay phenomenon, aligning with our theoretical expectations. As depicted in Figure 1a, the decreasing rate of the test error decelerates as N exceeds a certain threshold. This can be attributed to the fact that when N and T are sufficiently large, the bias term τ arising from the discretization becomes the dominant factor in the error.

Conclusion

The ubiquity of correlated data in processes modeled with spatially-inhomogeneous diffusions has created substantial barriers to analysis. In this paper, we construct and analyze a neural network-based numerical algorithm for estimating multidimensional spatially-inhomogeneous diffusion processes based on discretely-observed data obtained from a single trajectory. Utilizing β-mixing conditions and local Rademacher complexity arguments, we establish the convergence rate for our neural diffusion estimator. Our upper bound has recovered the minimax optimal nonparametric function estimation rate in the common i.i.d. setting, even with correlated data. We expect our proof techniques serve as a model for general exponential ergodic diffusion processes beyond the toroidal setting considered here. Numerical experiments validate our theoretical findings and demonstrate the potential of applying the neural diffusion estimators across various contexts with provable accuracy guarantees. Extending our results to typical biophysical settings, *e.g.* compact domains with reflective boundaries and motion blur due to measurement error, could help establish more rigorous error estimates for physical inference problems.

Acknowledgements

Grant M. Rotskoff acknowledges support from a Google Research Scholar Award.

References

Aeckerle-Willems, C.; and Strauch, C. 2018. Sup-norm adaptive simultaneous drift estimation for ergodic diffusions. arXiv:1808.10660.

Aït-Sahalia, Y. 2002. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, 70(1): 223–262.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497 – 1537.

Bibby, B. M.; and Sørensen, M. 1995. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 17–39.

Bogachev, V. I.; Krylov, N. V.; Röckner, M.; and Shaposhnikov, S. V. 2022. *Fokker–Planck–Kolmogorov Equations*, volume 207. American Mathematical Society.

Boucheron, S.; Lugosi, G.; and Massart, P. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bradley, R. C. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2(none): 107 – 144.

Chen, M.; Jiang, H.; Liao, W.; and Zhao, T. 2022. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4): 1203–1253.

Chen, Z.; Lu, J.; and Lu, Y. 2021. On the representation of solutions to elliptic pdes in barron spaces. *Advances in neural information processing systems*, 34: 6454–6465.

Crommelin, D.; and Vanden-Eijnden, E. 2011. Diffusion estimation from multiscale data by operator eigenpairs. *Multiscale Modeling & Simulation*, 9(4): 1588–1623.

Dalalyan, A. 2005. Sharp adaptive estimation of the drift function for ergodic diffusions. *The Annals of Statistics*, 33(6): 2507 – 2528.

Dalalyan, A.; and Reiß, M. 2006. Asymptotic statistical equivalence for scalar ergodic diffusions. *Probability theory and related fields*, 134: 248–282.

Dalalyan, A.; and Reiß, M. 2007. Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case. *Probability theory and related fields*, 137: 25–47.

Davis, W.; and Buffett, B. 2022. Estimation of drift and diffusion functions from unevenly sampled time-series data. *Physical Review E*, 106(1): 014140.

de la Peña, V. H.; Klass, M. J.; and Leung Lai, T. 2004. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3): 1902 – 1933.

Down, D.; Meyn, S. P.; and Tweedie, R. L. 1995. Exponential and uniform ergodicity of Markov processes. *The Annals of Probability*, 23(4): 1671–1691.

Duan, C.; Jiao, Y.; Lai, Y.; Lu, X.; and Yang, Z. 2021. Convergence rate analysis for deep ritz method. arXiv:2103.13330.

Elerian, O.; Chib, S.; and Shephard, N. 2001. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4): 959–993.

Frishman, A.; and Ronceray, P. 2020. Learning force fields from stochastic trajectories. *Physical Review X*, 10(2): 021009.

Gu, Y.; Harlim, J.; Liang, S.; and Yang, H. 2023. Stationary density estimation of itô diffusions using deep learning. *SIAM Journal on Numerical Analysis*, 61(1): 45–82.

Han, J.; Jentzen, A.; and E, W. 2018. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34): 8505–8510.

Hang, H.; Feng, Y.; Steinwart, I.; and Suykens, J. A. 2016. Learning theory estimates with observations from general stationary stochastic processes. *Neural computation*, 28(12): 2853–2889.

Hang, H.; and Steinwart, I. 2017. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2): 708 – 743.

Hoffmann, M. 1997. Minimax estimation of the diffusion coefficient through irregular samplings. *Statistics & probability letters*, 32(1): 11–24.

Hoffmann, M. 1999a. Adaptive estimation in diffusion processes. *Stochastic processes and their Applications*, 79(1): 135–163.

Hoffmann, M. 1999b. Lp estimation of the diffusion coefficient. *Bernoulli*, 447–481.

Hütter, J.-C.; and Rigollet, P. 2021. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2): 1166 – 1194.

Karniadakis, G. E.; Kevrekidis, I. G.; Lu, L.; Perdikaris, P.; Wang, S.; and Yang, L. 2021. Physics-informed machine learning. *Nature Reviews Physics*, 3(6): 422–440.

Khoo, Y.; Lu, J.; and Ying, L. 2021. Solving parametric PDE problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3): 421–435.

Koltchinskii, V. 2006. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6): 2593 – 2656.

Kovachki, N.; Li, Z.; Liu, B.; Azizzadenesheli, K.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2021. Neural operator: Learning maps between function spaces. arXiv:2108.08481.

Kulik, A. 2017. Ergodic behavior of Markov processes. In *Ergodic Behavior of Markov Processes*. de Gruyter.

Kuznetsov, V.; and Mohri, M. 2017. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1): 93–117.

Lau, A. W.; and Lubensky, T. C. 2007. State-dependent diffusion: Thermodynamic consistency and its path integral formulation. *Physical Review E*, 76(1): 011123.

Li, H.; Khoo, Y.; Ren, Y.; and Ying, L. 2022. A semigroup method for high dimensional committor functions based on neural network. In *Mathematical and Scientific Machine Learning*, 598–618. PMLR.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2021. Markov neural operators for learning chaotic systems. arXiv:2106.06898.

Lin, B.; Li, Q.; and Ren, W. 2023. Computing high-dimensional invariant distributions from noisy data. *Journal of Computational Physics*, 474: 111783.

Long, Z.; Lu, Y.; Ma, X.; and Dong, B. 2018. Pde-net: Learning pdes from data. In *International conference on machine learning*, 3208–3216. PMLR.

López-Pérez, A.; Febrero-Bande, M.; and González-Manteiga, W. 2021. Parametric estimation of diffusion processes: A review and comparative study. *Mathematics*, 9(8): 859.

Lu, L.; Jin, P.; and Karniadakis, G. E. 2019. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. arXiv:1910.03193.

Lu, Y.; Blanchet, J.; and Ying, L. 2022. Sobolev acceleration and statistical optimality for learning elliptic equations via gradient descent. *Advances in Neural Information Processing Systems*, 35: 33233–33247.

Lu, Y.; Chen, H.; Lu, J.; Ying, L.; and Blanchet, J. 2021. Machine learning for elliptic pdes: Fast rate generalization bound, neural scaling law and minimax optimality. *arXiv preprint arXiv:2110.06897*.

Marwah, T.; Lipton, Z.; and Risteski, A. 2021. Parametric complexity bounds for approximating PDEs with neural networks. *Advances in Neural Information Processing Systems*, 34: 15044–15055.

Marwah, T.; Lipton, Z. C.; Lu, J.; and Risteski, A. 2023. Neural network approximations of pdes beyond linearity: A representational perspective. In *International Conference on Machine Learning*, 24139–24172. PMLR.

Mohri, M.; and Rostamizadeh, A. 2008. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21.

Mohri, M.; and Rostamizadeh, A. 2010. Stability Bounds for Stationary φ-mixing and β-mixing Processes. *Journal of Machine Learning Research*, 11(2).

Nickl, R. 2022. Inference for diffusions from low frequency measurements. arXiv:2210.13008.

Nickl, R.; and Ray, K. 2020. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *The Annals of Statistics*, 48(3): 1383 – 1408.

Nickl, R.; and Söhl, J. 2017. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *The Annals of Statistics*, 45(4): 1664 – 1693.

Nickl, R.; van de Geer, S.; and Wang, S. 2020. Convergence rates for penalized least squares estimators in PDE constrained regression problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1): 374–413.

Oga, A.; and Koike, Y. 2023. Drift estimation for a multidimensional diffusion process using deep neural networks. *Stochastic Processes and their Applications*, 104240.

Oono, K.; and Suzuki, T. 2019. Approximation and non-parametric estimation of ResNet-type convolutional neural networks. In *International conference on machine learning*, 4922–4931. PMLR.

Ozaki, T. 1992. A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach. *Statistica Sinica*, 113–135.

Papanicolau, G.; Bensoussan, A.; and Lions, J.-L. 1978. *Asymptotic analysis for periodic structures*. Elsevier.

Papaspiliopoulos, O.; Pokern, Y.; Roberts, G. O.; and Stuart, A. M. 2012. Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, 99(3): 511–531.

Pedersen, A. R. 1995. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, 257–279.

Pokern, Y.; Stuart, A. M.; and van Zanten, J. H. 2013. Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stochastic Processes and their Applications*, 123(2): 603–628.

Pokern, Y.; Stuart, A. M.; and Vanden-Eijnden, E. 2009. Remarks on drift estimation for diffusion processes. *Multiscale modeling & simulation*, 8(1): 69–95.

Rotskoff, G. M.; Mitchell, A. R.; and Vanden-Eijnden, E. 2022. Active importance sampling for variational objectives dominated by rare events: Consequences for optimization and generalization. In *Mathematical and Scientific Machine Learning*, 757–780. PMLR.

Rotskoff, G. M.; and Vanden-Eijnden, E. 2019. Dynamical computation of the density of states and Bayes factors using nonequilibrium importance sampling. *Physical review letters*, 122(15): 150602.

Schmidt-Hieber, J. 2020. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4): 1875 – 1897.

Shoji, I.; and Ozaki, T. 1998. Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4): 733–752.

Sørensen, H. 2004. Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3): 337–354.

Steinwart, I.; and Christmann, A. 2009. Fast learning from non-iid observations. *Advances in neural information processing systems*, 22.

Strauch, C. 2015. Sharp adaptive drift estimation for ergodic diffusions: the multivariate case. *Stochastic Processes and their Applications*, 125(7): 2562–2602.

Strauch, C. 2016. Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions. *Probability Theory and Related Fields*, 164: 361–400.

Stroock, D. W.; and Varadhan, S. S. 1997. *Multidimensional diffusion processes*, volume 233. Springer Science & Business Media.

Sura, P.; and Barsugli, J. 2002. A note on estimating drift and diffusion parameters from timeseries. *Physics Letters A*, 305(5): 304–311.

Suzuki, T. 2018. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. arXiv:1810.08033.

Tsybakov, A. B.; and Zaiats, V. G. 2009. *Introduction to nonparametric estimation*. Springer-Verlag New York.

Van der Meulen, F.; Van Der Vaart, A. W.; and Van Zanten, J. 2006. Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli*, 12(5): 863–888.

van Waaij, J.; and van Zanten, H. 2016. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electronic Journal of Statistics*, 10(1): 628 – 645.

Wang, X.; Feng, J.; Liu, Q.; Li, Y.; and Xu, Y. 2022. Neural network-based parameter estimation of stochastic differential equations driven by Lévy noise. *Physica A: Statistical Mechanics and its Applications*, 606: 128146.

Weinan, E.; and Wojtowytsch, S. 2022. Some observations on high-dimensional partial differential equations with Barron data. In *Mathematical and Scientific Machine Learning*, 253–269. PMLR.

Xie, Z.; Kulasiri, D.; Samarasinghe, S.; and Rajanayaka, C. 2007. The estimation of parameters for stochastic differential equations using neural networks. *Inverse Problems in Science and Engineering*, 15(6): 629–641.

Yu, B. 1994. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 94–116.

Yu, B.; et al. 2018. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1): 1–12.

Zang, Y.; Bao, G.; Ye, X.; and Zhou, H. 2020. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411: 109409.

Zhang, L.; Han, J.; Wang, H.; Car, R.; and Weinan, E. 2018. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14): 143001.

Zhang, T. 2004. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1): 56–85.

Ziemann, I. M.; Sandberg, H.; and Matni, N. 2022. Single trajectory nonparametric learning of nonlinear dynamics. In *conference on Learning Theory*, 3333–3364. PMLR.