Error bounds of the invariant statistics in machine learning of ergodic Itô diffusions

He Zhang [a], John Harlim [a,b,*], Xiantao Li [a]

[a] *Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA*
[b] *Department of Meteorology and Atmospheric Science, Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA*

ABSTRACT

This paper studies the theoretical underpinnings of machine learning of ergodic Itô diffusions. The objective is to understand the convergence properties of the invariant statistics when the underlying system of stochastic differential equations (SDEs) is empirically estimated with a supervised regression framework. Using the perturbation theory of ergodic Markov chains and the linear response theory, we deduce a linear dependence of the errors of one-point and two-point invariant statistics on the error in the learning of the drift and diffusion coefficients. More importantly, our study shows that the usual L^2-norm characterization of the learning generalization error is insufficient for achieving this linear dependence result. We find that sufficient conditions for such a linear dependence result are through learning algorithms that produce a uniformly Lipschitz and consistent estimator in the hypothesis space that retains certain characteristics of the drift coefficients, such as the usual linear growth condition that guarantees the existence of solutions of the underlying SDEs. We examine these conditions on two well-understood learning algorithms: the kernel-based spectral regression method and the shallow random neural networks with the ReLU activation function.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Model error is inevitable, whether the model is formulated from direct empirical observations or deduced from fundamental physical principles, e.g., conservation laws. In this paper, we study modeling error arising from learning dynamical systems that obey a system of stochastic differential equations (SDEs) driven by Brownian noise [1–3], which are used in many scientific disciplines. In this context, the task is to identify the drift and diffusion coefficients from a time series of the SDEs. This inverse problem has been a central topic of interest for a long time and posted under various names, from parameter estimation, data-driven modeling, closure modeling, and lately, as a supervised learning task as machine learning becomes popular.

When the function forms are presumed, many classical methods, e.g., moment methods, maximum likelihood, and filtering, can be used [4]. Along this line, the MCMC-based Bayesian inference [5,6] is an important direction that allows for the estimation of the distribution of the parameters instead of point estimation in the traditional approaches. Since the same problem can be posed as a supervised learning task, a lot of recent interest has been shifting to machine learning approaches. Among the linear estimators, a popular approach is the kernel-based method [7–13], whose connection to the parametric modeling paradigm has been studied in [14]. In this direction, many nonparametric models have been proposed, including the orthogonal polynomials [10], wavelets [8], Gaussian processes [9], radial kernels [9], diffusion maps based models [15–17], just to name a few. Beyond the kernel approaches, the neural-network approach has been applied to estimate the drift coefficient [18] with application in biomolecular modeling, and the missing component in the drift term [19] with application to modeling atmospheric flow over topography.

Recurrent neural networks were shown to produce state-of-art numerical performances in learning high-dimensional nonlinear dynamical systems, even beyond SDEs [19–22]. These empirical successes, however, are not completely understood. Specifically, while the approximation theory of recurrent neural networks has been studied (see e.g., [23,24]), it remains unclear whether the neural

---

network model, obtained from a training procedure that involves solving a nonlinear, highly non-convex, optimization problem, can provide a convergence guarantee. Reservoir Computing (RC) [25] was introduced as an alternative to the tedious training procedure in the neural network model. This method, which is effectively a random neural network [26] in the context of recurrent neural network architecture, is based on the premise that fitting randomized function can be as effective yet computationally cheaper than solving the corresponding nonlinear optimization problem. This class of approach is effectively a conditionally linear estimator since it specifies the parameters in the activation function by randomly generated weights and trains the outer weights using the linear (ridge-) regression method. This surprisingly simple training procedure was shown to be effective in learning attractors of chaotic dynamical systems [27–29]. Recent theoretical results also shed some lights on its approximation and estimation properties [30,31], and universality in learning stochastic processes [32].

Building on the above independently reported positive successes, our goal is to understand under which conditions the underlying stochastic processes, driven by unknown SDEs, can be accurately emulated by a supervised learning procedure. Various metrics can be used to quantify the consistency of the estimated dynamics. For example, the (strong) pathwise error convergence that is classically used to characterize the numerical discretization error [1] has been used for quantifying the accuracy in learning partially known dynamics [19]. In the SDE application, they deduced under mild conditions that one can achieve accurate pathwise predictions up to a finite time with an error bound that is polynomial as a function of the learning error rate. In this paper, we will quantify the error in the estimation of one-point invariant statistics and two-point correlation statistics. Although these two metrics are commonly used to empirically assess the performance of the estimated dynamics through various algorithms [5,12,14,19,33–36], as a means to validate the consistency of the estimated dynamics, our emphasis is placed on the theoretical analysis. In particular, we will show that the errors in these statistics will depend linearly on a parameter ϵ that reflects the error in the estimation of drift and diffusion coefficients. This result not only guarantees the convergence of the invariant statistical estimation as $\epsilon \to 0$, but also provides a means for designing efficient learning algorithms when the parameter ϵ is specified as a function of the size of training data and other parameters that characterize the "size" of the hypothesis space, the strength of the noise, and the step size of the discrete-time series.

Our study will be based on the perturbation theory of ergodic Markov chains [37–39] and the linear response theory [40], which will be reviewed in Section 2. In the context of learning, we will specify the perturbation as the error induced by a regression learning framework in estimating the drift and diffusion coefficients of SDEs. Our main contribution, which will be discussed in Section 3, is to deduce error bounds of the estimation of one-point and two-point invariant statistics in terms of the error of the learning framework. Beyond these error bounds, the more important aspect of this study is to specify mathematical conditions that allow for the error bounds to be valid. Practically, these conditions allow one to pre-determine whether the proposed learning method is adequate or whether it can be adjusted to guarantee a convergent estimation. We will examine the validity of these conditions on two machine learning methods. In Section 4, we will discuss a kernel-based spectral regression method. We consider an RKHS induced by the orthonormal set of eigenfunctions of an integral operator defined over the invariant distribution of the data, which can be empirically estimated from the discrete samples. Subsequently, in Section 5, we will discuss a random neural network model of a simple single hidden-layer feed-forward neural network with ReLU activation function, which is a simple randomized function approximation relative to the reservoir computing. In these two sections, we will provide an overview of the generalization errors of these methods adopted in our application. This discussion is mainly based on the results in [30,41–44]. For these two machine learning algorithms, we will also examine the validity of the Assumption 3.1 that underpins the perturbation theory of Markov chains. In Section 6, we close the paper with a summary and some discussions on open issues.

2. Existing theory on statistics of perturbed Markov chains

In this section we will review the essential concepts and results in the perturbation theory for Markov chains [37,38] (Section 2.1), ergodic theory of SDEs [39] (Sections 2.2 and 2.3), and the long-time linear response theory [40] (Section 2.4). The theory involves both continuous Markov processes, e.g., Itô diffusions, and discrete Markov chains, e.g., the Euler–Maruyama approximation of the Itô diffusions. We will use the notation $\{\cdot(t)\}$, e.g., $\{X(t)\}$, and $\{\cdot_n\}$, e.g., $\{X_n\}$, to denote Markov processes and Markov chains, respectively. Throughout the paper, $\|\cdot\|$ always denotes the standard Euclidean norm in $\mathbb{R}^d$.

2.1. A perturbation theory for ergodic Markov chains

In this section, we review the perturbation theory for ergodic Markov chains in [37,38]. Let $\mathscr{B}(\mathbb{R}^d)$ denote the Borel σ-algebra on $\mathbb{R}^d$, and $\{u_n\}_{n=0}^{\infty}$ always denotes a Markov chain from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. In what follows, we will use the shorthand notation $|f| \leq V$ to mean $|f(x)| \leq V(x)$ for all $x \in \mathbb{R}^d$. Our first definition, following [38], focuses on the concept of geometrically ergodic Markov chains.

Definition 2.1. A Markov chain $\{u_n\}_{n=0}^{\infty}$ is *geometrically ergodic*, if:

   i. There exists a unique invariant measure, π, on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.
   ii. There exists a measurable function $V : \mathbb{R}^d \to [1, +\infty)$ such that

$$\mathbb{E}^x[V(u_n)] < \infty, \quad \forall n \geq 0,$$

   where $\mathbb{E}^x[\cdot]$ denotes the expectation under the initial condition $u_0 = x$.
   iii. Let $\mathscr{G}$ be the set of all measurable functions f with $|f| \leq V$. There exists a set $\mathscr{G}_0 \subset \mathscr{G}$ containing V such that

$$\sup_{f \in \mathscr{G}_0} \left| \mathbb{E}^x[f(u_n)] - \pi(f) \right| \leq R\rho^n V(x), \quad \forall n \geq 0, \tag{1}$$

   for some constants $R \in (0, +\infty)$ and $\rho \in (0, 1)$. Here, $\pi(f) := \int f \pi(\mathrm{d}x)$.

The inequality in (1), as the key component of the geometrical ergodicity, describes the decay rate of the V-norm distance [37] between the distribution of u_n and the invariant measure π as $n \to +\infty$. In general, we may consider other probability distances to characterize the convergence in (1), e.g., the Wasserstein distance [37]. Here, the set $\mathscr{G}_0$ is introduced to rule out certain "ill-behaved" observables for simplicity. In our later discussions, $\mathscr{G}_0$ is either $\mathscr{G}$ itself, e.g., in Theorem 2.1, or a set of locally Lipschitz functions, e.g., in Proposition 2.2.

We approximate the geometrically ergodic Markov chain $\{u_n\}$ by another perturbed Markov chain $\{u_n^\epsilon\}$ (which may not be geometrically ergodic), where ϵ represents the "scale" of the perturbation. To specify how the approximation error is carried over to the error of the resulting statistics in the long run, we will state the following perturbation bound, which is a result of the Corollary 3.4 in [37].

Proposition 2.1. *Let $\{u_n\}$ be a geometrically ergodic Markov chain satisfying the condition in Definition 2.1. We further assume that the Lyapunov function V satisfies*

$$\mathbb{E}^x[V(u_1)] \leq \alpha V(x) + \beta, \quad \forall x \in \mathbb{R}^d, \tag{2}$$

with constants $\alpha \in (0, 1)$ and $\beta \in (0, +\infty)$. Let $\{u_n^\epsilon\}$ be a perturbed Markov chain with respect to $\{u_n\}$. We define

$$\gamma := \sup_{x \in \mathbb{R}^d} \sup_{f \in \mathscr{G}_0} \frac{\left|\mathbb{E}^x[f(u_1^\epsilon)] - \mathbb{E}^x[f(u_1)]\right|}{V(x)}. \tag{3}$$

If $\gamma \in (0, 1 - \alpha)$, then, for any fixed initial condition $u_0 = u_0^\epsilon = x \in \mathbb{R}^d$, we have,

$$\sup_{f \in \mathscr{G}_0} \left|\mathbb{E}^x[f(u_n^\epsilon)] - \mathbb{E}^x[f(u_n)]\right| \leq R(1 - \rho^n)\frac{\gamma \kappa}{1 - \rho}, \quad \kappa := \max\left\{V(x), \frac{\beta}{1 - \gamma - \alpha}\right\}, \quad \forall n \geq 0, \tag{4}$$

for some constant $R \in (0, +\infty)$, where $\rho \in (0, 1)$ is defined by (1).

We would like to point out that the original error bound presented by Corollary 3.4 in [37] allows for u_n and u_n^ϵ to have different initial conditions.

As a direct consequence of Proposition 2.1 and Eq. (1), we have

$$\sup_{f \in \mathscr{G}_0} \left|\mathbb{E}^x[f(u_n^\epsilon)] - \pi(f)\right| \leq R\left[(1 - \rho^n)\frac{\gamma \kappa}{1 - \rho} + \rho^n V(x)\right], \quad \forall n \geq 0, \tag{5}$$

for some constant $R \in (0, +\infty)$. The fact that the error bound in (4) depends on γ, defined through the V-norm in (3), provides a convenient way for characterizing the errors of the estimated transition kernel. Namely, we only need to study the "one-step" error between the statistics of $\{u_1\}$ and $\{u_1^\epsilon\}$. When $\mathscr{G}_0 = \mathscr{G}$, the constant γ in (3) is also used as the upper bound of the V-norm difference between the transition kernels of $\{u_n\}$ and $\{u_n^\epsilon\}$ over $x \in \mathbb{R}^d$ [37].

2.2. The Itô diffusion and its approximations

The key results developed in our paper focus on the case where the underlying dynamical system is an Itô diffusion. In this section, we review some basic concepts and results related to the Itô diffusion and its approximations. For classical theory of Itô diffusions, readers may refer to [1–3].

A d-dimensional Itô diffusion is a SDE of the form

$$\dot{X}(t) = b(X(t)) + \sigma(X(t))\dot{W}, \quad X(0) = x, \quad t \geq 0, \tag{6}$$

where $x \in \mathbb{R}^d$, $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d \times m}$ ($m \leq d$) are the initial condition, drift and diffusion coefficients, respectively. The process W in (6) denotes a standard m-dimensional Brownian motion. When $m < d$, the noise in (6) is degenerate. Itô diffusions with degenerate noise are common in applications, e.g., Langevin dynamics [3]. Non-degeneracy often makes it convenient to prove certain properties of the corresponding Itô diffusions, e.g., the ergodicity [39] and the regularity of the invariant measure [45]. But for the sake of generality, in our paper, we will consider Itô diffusions with possible degenerate noise. We propose the following assumption on the coefficients.

Assumption 2.1. The coefficients b and σ in (6) are Borel measurable and satisfy the following conditions:

i. **Globally Lipschitz condition:** There exists a constant $K_1 \in (0, +\infty)$ such that

$$\|b(x) - b(y)\| \leq K_1\|x - y\|, \quad \|\sigma(x) - \sigma(y)\|_F \leq K_1\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

ii. **Linear growth bound:** There exists a constant $K_2 \in (0, +\infty)$ such that

$$\|b(x)\|^2 \leq K_2^2(1 + \|x\|^2), \quad \|\sigma(x)\|_F^2 \leq K_2^2(1 + \|x\|^2), \quad \forall x, y \in \mathbb{R}^d.$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm, that is, $\|\sigma\|_F = \left(\sum_{i,j} \sigma_{ij}^2\right)^{\frac{1}{2}}$.

We shall henceforth hold fixed a Brownian motion W, and the associated family of σ-algebra $\{\mathscr{A}_t, t \geq 0\}$. Assumption 2.1 (i) ensures the existence and uniqueness of the $\{\mathscr{A}_t\}$-adapted strong solution of (6) [1,2]. In particular, the following lemma provides useful bounds on the even order moments of the solution to (6).

Lemma 2.1. *Suppose Assumption 2.1 (ii). Then, for any integer $p \geq 1$, the solution $X(t)$ of (6) satisfies*

$$\mathbb{E}^x\left[\|X(t)\|^{2p}\right] \leq e^{2K_2(4p^2+2p)t}(1 + \|x\|^{2p}), \quad \forall t \geq 0,$$

where the constant K_2 is the same as in Assumption 2.1.

Proof. The proof for $d = 1$ can be found in [1] (Theorem 4.5.4), which can be directly generalized to higher dimensional cases. □

The Itô diffusion (6) defines a time-continuous Markov process. In practice, we may introduce the corresponding Markov chains either by sampling or numerical discretizations. By sampling, we mean the resulting Markov chain $\{X_n\}$ satisfies $X_n = X(n\delta)$, for some $\delta > 0$. For numerical discretizations, a classical example is the Euler–Maruyama (EM) scheme [1].

For a fixed step size $\delta > 0$, the EM scheme generates a Markov chain $\{X_n^\delta\}$ via,

$$X_0^\delta = x, \quad X_{n+1}^\delta = X_n^\delta + \delta b(X_n^\delta) + \sqrt{\delta}\sigma(X_n^\delta)\xi_n, \quad n = 0, 1, \ldots, \tag{7}$$

where $\{\xi_n\}$ denotes a sequence of independent, identically distributed, m-dimensional standard Gaussian random variables. The EM scheme attains the $1/2$-order of strong convergence [1], that is, under Assumption 2.1, there exist constants $R, D \in (0, +\infty)$ such that

$$\mathbb{E}^x\left[\|X_n - X_n^\delta\|^2\right] \leq Re^{Dn\delta}(1 + \|x\|^2)\delta, \quad n = 0, 1, \ldots, \tag{8}$$

where the constants $R, D \in (0, +\infty)$ are independent of δ. Moreover, X_n^δ yields similar moment bounds as in Lemma 2.1 (see Theorem 10.2.2 in [1] for details). It is worthwhile to mention that the bounds discussed in Lemma 2.1 and Eq. (8) can be improved under extra assumptions, e.g., drift coefficients b satisfying dissipative conditions [46].

2.3. The ergodic Itô diffusions

In Section 2.1, we have discussed a perturbation result for ergodic Markov chains. In this paper, the Markov chains are generated from ergodic Itô diffusions by sampling or numerical discretization. Thus, we need to first inspect the concept of ergodic Itô diffusions and how they are connected to ergodic Markov chains. The unlisted proofs in this section can be found in [39]. Following [39], we consider Itô diffusions (6) with additive noise,

$$\dot{X}(t) = b(X(t)) + \sigma\dot{W}, \quad X(0) = x, \tag{9}$$

where $\sigma \in \mathbb{R}^{d \times m}$ ($m \leq d$). The fixed constant matrix σ is assumed to have linearly independent column vectors, that is, $\text{rank}(\sigma) = m$. Here, $\{X(t)\}$ in (9) forms a Markov process on the state space $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. We denote the transition kernel of the Markov process $\{X(t)\}$ by

$$P_t(x, A) := \mathbb{P}(X(t) \in A \mid X(0) = x), \quad t \geq 0, \ x \in \mathbb{R}^d, \ A \in \mathscr{B}(\mathbb{R}^d). \tag{10}$$

To establish the geometric ergodicity for the system (9), we introduce the following series of assumptions [39].

Assumption 2.2. The transition kernel P_t in (10) satisfies, for some fixed compact set $S \in \mathscr{B}(\mathbb{R}^d)$, the following:

   i. For some $y^* \in \text{int}(S)$ (the interior of S) and any $r > 0$, there is a $t_1 = t_1(r) > 0$ such that

   $$P_{t_1}(x, \mathscr{B}_r(y^*)) > 0, \quad \forall x \in S.$$

   ii. For any $t > 0$ the transition kernel yields a density $p_t(x, y)$, that is,

   $$P_t(x, A) = \int_A p_t(x, y)\mathrm{d}y, \quad \forall x \in S, \ A \in \mathscr{B}(\mathbb{R}^d) \cap \mathscr{B}(S),$$

   and $p_t(x, y)$ is jointly continuous in $(x, y) \in S \times S$.

Here, $\mathscr{B}_r(y^*)$ denotes the open ball of radius r centered at y^*, and $\mathscr{B}(S)$ denotes the sub-σ-algebra on S with respect to $\mathscr{B}(\mathbb{R}^d)$.

Assumption 2.3. There is a function $V : \mathbb{R}^d \to [1, +\infty)$, with $\lim_{x\to\infty} V(x) = +\infty$, and $a_1, d_1 \in (0, +\infty)$ such that

$$\mathscr{L}V(x) \leq -a_1 V(x) + d_1, \quad \forall x \in \mathbb{R}^d.$$

Here $\mathscr{L}$ is the generator for (9) given by

$$\mathscr{L}f = \sum_{i=1}^d b_i \frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^d (\sigma\sigma^\top)_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j},$$

where $(\sigma\sigma^\top)_{ij}$ denotes the ij-component of the matrix $\sigma\sigma^\top$.

As a direct consequence of Assumption 2.3 and the Dynkin's formula [2,39], we have

$$\mathbb{E}^x[V(X(t))] \leq e^{-a_1 t}V(x) + \frac{d_1}{a_1}(1 - e^{-a_1 t}), \quad \forall x \in \mathbb{R}^d, \ \forall t \geq 0,$$

where $\mathbb{E}^x[\cdot]$ denotes the expectation under (9), with respect to the initial condition $X_0 = x$. When $t = \delta$, we have

$$\mathbb{E}^x[V(X_1)] = \mathbb{E}^x[V(X(\delta))] \leq e^{-a_1\delta}V(x) + \frac{d_1}{a_1}(1 - e^{-a_1\delta}), \quad \forall x \in \mathbb{R}^d,$$

that is, the Markov chain $\{X_n = X(n\delta)\}$ satisfies the condition (2) in Proposition 2.1. The following theorem (Theorem 2.5 in [39]), guaranteeing the geometric ergodicity of the Markov chain $\{X_n\}$, is the foundation of our study of the perturbation theory.

Theorem 2.1. *Let $\{X(t)\}$ be the Markov process defined in* (9) *that satisfies Assumptions* 2.2 *and* 2.3 *with the compact set S given by*

$$S = \left\{ x \in \mathbb{R}^d \mid V(x) \leq \frac{2d_1}{a_1(\zeta - e^{-a_1\delta})} \right\},$$

for some $\zeta \in (e^{-a_1\delta/2}, 1)$ and $\delta > 0$. Then there exists a unique invariant measure π. Furthermore, there exist $\rho = \rho(\zeta) \in (0, 1)$ and $R = R(\zeta) \in (0, +\infty)$ such that

$$\sup_{f \in \mathscr{G}} \left| \mathbb{E}^x[f(X_n)] - \pi(f) \right| \leq R\rho^n V(x), \quad \forall x \in \mathbb{R}^d, \ \forall n \geq 0,$$

where $\mathscr{G}$ denotes the set of all measurable functions with $|f| \leq V$, that is, $\{X_n\}$ is a geometrically ergodic Markov chain defined in Definition 2.1 *with $\mathscr{G}_0 = \mathscr{G}$.*

In [39], Theorem 2.1 has been applied to a variety of SDEs, including the Langevin dynamics, monotone and dissipative systems, and stochastic gradient systems. The function V in Assumption 2.3 is called the *Lyapunov function* of the dynamical system (9). In particular, we further assume that V is of a polynomial growth rate.

Assumption 2.4. The Lyapunov function V in Assumption 2.3 is of the form $V = W^\ell$ for some $\ell \geq 1$, where W is essentially quadratic, i.e., there exist constants $C_i \in (0, +\infty)$, $i = 1, 2, 3$, such that

$$C_1 \left(1 + \|x\|^2\right) \leq W(x) \leq C_2 \left(1 + \|x\|^2\right), \quad \|\nabla W(x)\| \leq C_3 \left(1 + \|x\|\right), \quad \forall x \in \mathbb{R}^d. \tag{11}$$

Assumption 2.4 is not only useful in deriving perturbation bounds in Section 3 (see Lemma 3.1 for the details), it also ensures that any "reasonable" numerical discretization scheme will inherit the ergodicity of (9) [39]. In particular, we have the following proposition.

Proposition 2.2. *Let* Assumptions 2.1–2.4 *hold. Then, there exists $\delta_0 > 0$, such that $\forall \delta \in (0, \delta_0)$ the Markov chain generated by the EM scheme with step size δ, $\{X_n^\delta\}$ in* (7)*, is geometrically ergodic with invariant measure $\tilde{\pi}^\delta$ and with same Lyapunov function $V(x)$ as in* Assumption 2.3*. In particular, we define*

$$\mathscr{G}_\ell := \left\{ f \in \mathscr{G} \mid |f(x) - f(y)| \leq C_\ell \left(1 + \|x\|^{2\ell-1} + \|y\|^{2\ell-1}\right) \|x - y\|, \quad \forall x, y \in \mathbb{R}^d \right\},$$

as the set "$\mathscr{G}_0$" (in Definition 2.1*) for $\{X_n^\delta\}$, where $C_\ell > 0$ is a fixed constant such that $\mathscr{G}_\ell$ contains $V(x)$. We have the following results:*

i. *There exists $a_2 = a_2(\delta) \in (0, a_1)$ (a_1 and d_1 are defined in* Assumption 2.3*) such that*

$$\mathbb{E}^x[V(X_1^\delta)] \leq e^{-a_2\delta} V(x) + \frac{d_1}{a_2}, \quad \forall x \in \mathbb{R}^d.$$

ii. *There exist $R = R(\ell, \delta) \in (0, +\infty)$ and $D = D(\ell, \delta) \in (0, +\infty)$ such that,*

$$\sup_{f \in \mathscr{G}_\ell} \left| \mathbb{E}^x[f(X_n^\delta)] - \tilde{\pi}^\delta(f) \right| \leq Re^{-Dn\delta} V(x), \quad \forall x \in \mathbb{R}^d, \ \forall n \geq 0.$$

iii. *There exist $K = K(\ell)$ and $\nu \in (0, 1/2)$ independent of ℓ, such that*

$$\sup_{f \in \mathscr{G}_\ell} \left| \pi(f) - \tilde{\pi}^\delta(f) \right| \leq K\delta^\nu \pi(V). \tag{12}$$

Here, the set $\mathscr{G}_\ell$ is well-defined since by Assumption 2.4 and Eq. (11), one has,

$$\|\nabla V(x)\| = \ell W^{\ell-1} \|\nabla W(x)\| \leq \ell C_2^{\ell-1} C_3 (1 + \|x\|^2)^{\ell-1}(1 + \|x\|),$$

which leads to

$$|V(x) - V(y)| \leq \int_0^1 \|\nabla V(sx + (1-s)y)\| \|x - y\| \mathrm{d}s \leq \ell C_2^{\ell-1} C_3(1 + \|x\|^2 + \|y\|^2)^{\ell-1}(1 + \|x\| + \|y\|)\|x - y\|$$

$$\leq C_\ell(1 + \|x\|^{2\ell-1} + \|y\|^{2\ell-1})\|x - y\|,$$

for some constant $C_\ell = O(2^{2\ell})$. Proposition 2.2 is a direct corollary of Theorem 7.3 in [39], whose proof is closely related to the result in [38].

2.4. The long-time linear response theory

So far, all the perturbation bounds introduced only focus on the one-point statistics given an observable $f \in \mathscr{G}_0$ satisfying $|f| \leq V$. However, the set of admissible observables $\mathscr{G}_0$ is not general enough for our implementation. As a remedy, we will review the long-time linear response theory, which is justified in [40] in an abstract setting. This will help us capture the leading order term of the error for more general observables.

We consider a family of Markov evolution operators $\{\mathscr{P}_t^\epsilon \mid t \geq 0, \ \epsilon \in (-\epsilon_0, \epsilon_0)\}$ on $\mathbb{R}^d$ that characterize the unperturbed dynamics in (9) and its perturbations. We will specify such $\mathscr{P}_t^\epsilon$ in Section 3. Here, to help readers understand the notations, one can interpret the parameter ϵ as the strength of the perturbation. In other words, when $\epsilon = 0$, $\mathscr{P}_t^0$ reduces to the evolution operator of the unperturbed dynamics, e.g., the Itô diffusion (9). Namely,

$$\left(\mathscr{P}_t^0 f\right)(x) = \int f(y)P_t(x, \mathrm{d}y) = \mathbb{E}^x[f(X(t))],$$

where the transition kernel P_t is defined in (10). We are interested in the long-time behavior of the perturbed system described by $\mathscr{P}_t^\epsilon$ for ϵ close to 0, which requires the following assumption [40].

Assumption 2.5. There exists an $\epsilon_0 > 0$ such that for all $\epsilon \in (-\epsilon_0, \epsilon_0)$, $\mathscr{P}_t^\epsilon$ yields an invariant probability measure π^ϵ on $\mathbb{R}^d$.

When the unperturbed dynamics corresponds to an ergodic Itô diffusion (9), by Theorem 2.1, we know there exists a unique invariant measure π for $\mathscr{P}_t^0$, that is, $\pi^0 = \pi$. But for $\epsilon \neq 0$, Assumption 2.5 only ensures the existence of the invariant measure. The aim of the long-time linear response theory is to show that the map

$$\epsilon \mapsto \pi^\epsilon(f)$$

is differentiable at $\epsilon = 0$ for every sufficiently regular observable $f : \mathbb{R}^d \to \mathbb{R}$. We will briefly review the assumptions in [40] that lead to the desirable result.

Let $C_c^\infty(\mathbb{R}^d)$ be the set of all smooth functions $f : \mathbb{R}^d \to \mathbb{R}$ that are compactly supported. Given continuous functions $G, H, U : \mathbb{R}^d \to [1, +\infty)$, we set $C_{G,H}^1$ to be the closure of $C_c^\infty(\mathbb{R}^d)$ under the norm

$$\|f\|_{1;G,H} := \sup_{x \in \mathbb{R}^d} \left(\frac{|f(x)|}{G(x)} + \frac{\|\nabla f(x)\|}{H(x)} \right), \tag{13}$$

and C_U to be the weighted space of continuous functions obtained by completing $C_c^\infty(\mathbb{R}^d)$ under the norm

$$\|f\|_U = \sup_{x \in \mathbb{R}^d} \frac{|f(x)|}{U(x)}.$$

The following assumption targets the spectral gap of $\mathscr{P}_t^0$ as an operator on $C_{G,H}^1$.

Assumption 2.6. There exist a time $t > 0$ and a constant $\lambda \in (0, 1)$ such that

$$\|\mathscr{P}_t^0 f - \pi(f)\|_{1;G,H} \leq \lambda \|f - \pi(f)\|_{1;G,H}, \quad \forall f \in C_{G,H}^1.$$

Assumption 2.6 also implies that the invariant measure, π, of the unperturbed dynamics (9) is unique [40], that is, $\text{Null}(I - (\mathscr{P}_t^0)^*) = \text{span}\{\pi\}$, where $(\mathscr{P}_t^0)^*$ denotes the adjoint operator of $\mathscr{P}_t^0$. Therefore, by Fredholm theorem, we have $\text{Range}(I - \mathscr{P}_t^0) = \text{Null}(I - (\mathscr{P}_t^0)^*)^\perp$, which means for every function $\varphi \in C_{G,H}^1$ centered with respect to π, there exists a unique function $\psi \in C_{G,H}^1$ such that

$$\psi - \mathscr{P}_t^0 \psi = \varphi,$$

and ψ is also centered with respect to π. We will henceforth use the notation $\psi = (I - \mathscr{P}_t^0)^{-1}\varphi$. Our next assumption concerns the Fréchet derivative of $\mathscr{P}_t^\epsilon$ with respect to ϵ.

Assumption 2.7. Let $C_{G,H}^1$ be the same as in Assumption 2.6. There exists a continuous function $U \geq G$ such that, for some fixed $t > 0$ and every $f \in C_{G,H}^1$, the map $\epsilon \mapsto \mathscr{P}_t^\epsilon f$, viewed as a map from $(-\epsilon_0, \epsilon_0)$ to C_U, is differentiable on $(-\epsilon_0, \epsilon_0)$. Denoting this Fréchet derivative by $\partial \mathscr{P}_t^\epsilon$, we furthermore assume that,

$$\left\| \partial \mathscr{P}_t^0 f \right\|_U \leq C \|f\|_{1;G,H}, \quad \forall f \in C_{G,H}^1,$$

for some constant $C \in (0, +\infty)$ independent of f.

Finally, we assume that we have an a priori bound on the integrability of the invariant measures.

Assumption 2.8. For U in Assumption 2.7 and π^ϵ in Assumption 2.5, we have

$$\sup_{\epsilon \in (-\epsilon_0, \epsilon_0)} \pi^\epsilon(U) = \sup_{\epsilon \in (-\epsilon_0, \epsilon_0)} \int U(x) \pi^\epsilon(\mathrm{d}x) < \infty.$$

Assumption 2.8 ensures that observables in C_U yield finite first moments with respect to the invariant measures π^ϵ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$. We state the following theorem in [40].

Theorem 2.2. Let $\{\mathscr{P}_t^\epsilon \mid \epsilon \in (-\epsilon_0, \epsilon_0)\}$ be a family of Markov evolution operators over $\mathbb{R}^d$ such that there exist C^1 functions $U, G, H : \mathbb{R}^d \to [1, +\infty)$ such that Assumptions 2.5–2.8 hold for some fixed $t > 0$. Then, the map $\epsilon \to \pi^\epsilon(f)$ is differentiable at $\epsilon = 0$ for all $f \in C_{G,H}^1$. In particular, we have

$$\frac{d}{d\epsilon} \pi^\epsilon(f) \Big|_{\epsilon=0} = \mathbb{E}_\pi \left[\partial \mathscr{P}_t^0 (I - \mathscr{P}_t^0)^{-1} (f - \pi(f)) \right], \tag{14}$$

where the right-hand side, as an expectation with respect to the invariant measure $\pi = \pi^0$, is well-defined.

Using Theorem 2.2, we can capture the leading order term of the error $|\pi^\epsilon(f) - \pi(f)|$ for $f \in C_{G;H}^1$. Although the result in Theorem 2.2 is observable-dependent, unlike the error bounds reviewed in Section 2.1, where the inequalities act as uniform bounds for a class of observables, Eq. (14) can be applied to more general observables, e.g., observables not controlled by the Lyapunov function V. To some extent, we trade the uniformity for generality.

Before we close the section, we would like to point out that the linear response theory can also be applied to the short-time response of the dynamics subject to perturbations [3]. We list Theorem 2.2 in this paper to prove Proposition 3.2 in Section 3.3.

3. Error bounds of the invariant statistics in learning ergodic Itô diffusions

In Section 2, we have reviewed the concept of geometrically ergodic Markov chain and discussed a series of results and bounds for ergodic Itô diffusions in a relatively abstract setting. In this section, we reformulate the perturbation theory as a problem in the context of learning the dynamical system (9). In particular, we will specify the perturbation as the error induced by the statistical learning (Section 3.1) and develop error bounds for both one-point statistics (Section 3.2) and two-point statistics (Section 3.3) based on the results in Section 2.

3.1. Learning ergodic Itô diffusions

Interpreting the system (9) as the unperturbed dynamics, under the same initial condition, we introduce a family of perturbed dynamics of the form,

$$\dot{X}^\epsilon(t) = b_\epsilon(X^\epsilon(t)) + \sigma_\epsilon \dot{W}, \quad X_0^\epsilon = x, \quad 0 < \epsilon \ll 1, \tag{15}$$

where $b_\epsilon : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma_\epsilon \in \mathbb{R}^{d \times m}$, the perturbed drift and diffusion coefficients, respectively, are parameterized by a parameter $\epsilon > 0$ corresponding to the "scale" of the perturbation. For simplicity, we assume W in (15) to be the same standard m-dimensional Brownian motion as in the unperturbed dynamics (9). We should point out that since we are interested in the error bound of the invariant statistics (rather than the pathwise error between $X(t)$ and $X^\epsilon(t)$), the perturbed diffusion coefficients σ_ϵ is defined so that $\sigma_\epsilon \sigma_\epsilon^\top$ is an estimate of $\sigma \sigma^\top$. To gain an intuition of the perturbed dynamics in (15) and develop a proper interpretation of the parameter ϵ, we introduce the following regression problem in estimating the drift coefficient b of the unperturbed dynamics (9).

Consider the Markov chain, $\{X_n^\delta\}$, generated by EM discretization of the unperturbed dynamics (9). Based on the numerical scheme in (7), we define the finite difference process,

$$Y_n^\delta := \frac{1}{\delta}\left(X_{n+1}^\delta - X_n^\delta\right) = b(X_n^\delta) + \delta^{-\frac{1}{2}}\sigma\xi_n. \tag{16}$$

Since ξ_n is independent of X_n^δ, we may express the drift coefficient b as the following conditional expectation,

$$b(x) = \mathbb{E}\left[Y_n^\delta \mid X_n^\delta = x\right], \quad \forall n \geq 0. \tag{17}$$

Let μ^δ denote the joint stationary distribution of the random variable $(X, Y) := (X_n^\delta, Y_n^\delta)$. Eq. (17) suggests that the solution to the following regression problem

$$\min_{h=(h_1,\dots,h_d)^\top} \mathcal{E}[h], \quad \mathcal{E}[h] := \mathbb{E}_{\mu^\delta}\left[\|h(X) - Y\|^2\right], \quad h_i \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta), \quad i = 1, 2\dots, d, \tag{18}$$

is an unbiased estimator of b [44]. Furthermore, the covariance matrix of the residual error satisfies

$$\mathbb{E}_{\mu^\delta}\left[(b(X) - Y)(b(X) - Y)^\top\right] = \delta^{-1}\sigma\sigma^\top.$$

Since the noise has independent components, the residual error is given by $\mathcal{E}[b] = \delta^{-1}\mathrm{Tr}[\sigma\sigma^\top]$, where $\mathrm{Tr}[\cdot]$ denotes the standard matrix trace operation. While this estimator is unbiased, the bias (also known as the approximation error [44]) may appear depending on the choice of the hypothesis space H, which will be clarified in Sections 4–5.

In practice, the loss function $\mathcal{E}[h]$ in (18) is approximated by an empirical loss function,

$$\mathcal{E}_N[h] := \frac{1}{N}\sum_{i=1}^N \|h(x_i) - y_i\|^2, \tag{19}$$

from i.i.d. samples $\{x_i, y_i\}_{i=1}^N$ of (X, Y) with stationary distribution μ^δ. Here, we consider i.i.d. samples only for the convenience of the theoretical analysis in Sections 4 and 5. Practically, the samples can be obtained by subsampling from the labeled time series $\{x_n^\delta, y_n^\delta\}_{n\geq 0}$ to reduce the temporal correlation, and thus the sampling error.

We define b_ϵ and $\sigma_\epsilon\sigma_\epsilon^\top$ as follows,

$$b_\epsilon := \arg\min_{h\in H} \mathcal{E}_N[h], \quad \sigma_\epsilon\sigma_\epsilon^\top := \frac{\delta}{N}\sum_{i=1}^N \left[(y_i - b_\epsilon(x_i))(y_i - b_\epsilon(x_i))^\top\right]. \tag{20}$$

One can see that the sample covariance is a biased estimator of $\sigma\sigma^\top$, that is,

$$\mathbb{E}_{\mu^\delta}[\sigma_\epsilon\sigma_\epsilon^\top] - \sigma\sigma^\top = \delta\mathbb{E}_{\tilde{\pi}^\delta}[(b(X) - b_\epsilon(X))(b(X) - b_\epsilon(X))^\top],$$

where, by Jensen's inequality, the bias satisfies

$$\left\|\mathbb{E}_{\tilde{\pi}^\delta}\left[(b(X) - b_\epsilon(X))(b(X) - b_\epsilon(X))^\top\right]\right\|_F \leq \mathbb{E}_{\tilde{\pi}^\delta}\left[\left\|(b(X) - b_\epsilon(X))(b(X) - b_\epsilon(X))^\top\right\|_F\right] = \mathbb{E}_{\tilde{\pi}^\delta}\left[\|(b(X) - b_\epsilon(X))\|^2\right]. \tag{21}$$

To have a better understanding of the error between $\sigma_\epsilon\sigma_\epsilon^\top$ and $\sigma\sigma^\top$, we introduce

$$D_i := \delta(y_i - b_\epsilon(x_i))(y_i - b_\epsilon(x_i))^\top - \sigma\sigma^\top - \delta\mathbb{E}_{\tilde{\pi}^\delta}[(b(X) - b_\epsilon(X))(b(X) - b_\epsilon(X))^\top], \quad i = 1, 2, \dots, N, \tag{22}$$

which defines a finite sequence of independent, random, symmetric matrices of mean 0. In particular, subtracting $\sigma_\epsilon\sigma_\epsilon^\top$ (20) from $\sigma\sigma^\top$ and using the definition in (22), we deduce that,

$$\left\|\sigma\sigma^\top - \sigma_\epsilon\sigma_\epsilon^\top\right\|_2 = \left\|\frac{1}{N}\sum_{i=1}^N D_i + \delta\mathbb{E}_{\tilde{\pi}^\delta}\left[(b(X) - b_\epsilon(X))(b(X) - b_\epsilon(X))^\top\right]\right\|_2$$

$$\leq \left\| \frac{1}{N} \sum_{i=1}^{N} D_i \right\|_2 + \delta \left\| \mathbb{E}_{\tilde{\pi}^\delta} \left[(b(X) - b_\epsilon(X))(b(X) - b_\epsilon(X))^\top \right] \right\|_2$$

$$\leq \left\| \frac{1}{N} \sum_{i=1}^{N} D_i \right\|_2 + \delta \mathbb{E}_{\tilde{\pi}^\delta} \left[\|(b(X) - b_\epsilon(X))\|^2 \right], \tag{23}$$

where $\| \cdot \|_2$ denotes the matrix 2-norm. Here, we have used the relation in (21) and the fact that $\|A\|_2 \leq \|A\|_F$ for any matrix A. The first term on the right-hand side of Eq. (23) is the error induced by the empirical estimation, can be bounded by the matrix Bernstein inequality, e.g., Theorem 6.2 in [47], assuming that $\tilde{\pi}^\delta$ belongs to the sub-exponential class. Ignoring the parameter δ, we will refer the second term as the generalization error of the learning algorithm of b. We will provide detailed discussions of this error term in Sections 4–5 for specific learning methods. Based on these observations, we define the parameter ϵ, which corresponds to the "scale" of the perturbation, as the spectral error of the diffusion matrix estimator, $\sigma_\epsilon \sigma_\epsilon^\top$,

$$\epsilon := \|\sigma \sigma^\top - \sigma_\epsilon \sigma_\epsilon^\top\|_2, \tag{24}$$

which is well-defined and is small in high probability for large enough N.

Our goal is to analyze the error bounds of the invariant statistics when the underlying ergodic Itô diffusion (9) is approximated by the perturbed dynamics in (15). In particular, we would like to understand how the error in the invariant statistics depends on ϵ in (24) and the generalization error, $\mathbb{E}_{\tilde{\pi}^\delta} \left[\|(b(X) - b_\epsilon(X))\|^2 \right]$. To develop results toward this direction, we need the following critical assumptions on the family of coefficients $\{b_\epsilon\}$ in (15) and their errors $\{b_\epsilon - b\}$. These assumptions not only elucidate the dependence on ϵ but also are conditions that we need to implement the results reviewed in Section 2.

Assumption 3.1. Consider the unperturbed and the family of perturbed Itô diffusions in (9) and (15), respectively. For any $0 < \epsilon \ll 1$ (ϵ defined in (24)), we assume

   i. The coefficient b_ϵ is Borel measurable and satisfies the globally Lipschitz condition as in Assumption 2.1 with uniform Lipschitz constant with respect to ϵ.

   ii. The family of coefficients $\{b_\epsilon\}$ is a sequence of consistent estimators of b in the space of continuous (vector-valued) functions of linear growth. That is,

$$\|b_\epsilon(x) - b(x)\|^2 \leq K_3^2 (1 + \|x\|^2)\epsilon^2, \quad \forall x \in \mathbb{R}^d, \tag{25}$$

     for some constant $K_3 \in (0, +\infty)$ independent of ϵ.

Assumption 3.1 (i) ensures that the coefficients b_ϵ in the perturbed dynamics (15) satisfy Assumption 2.1 with related constants, including the Lipschitz constants and those constants in the linear growth bound, independent of ϵ. The condition (25) suggests that not only $b_\epsilon \to b$ as $\epsilon \to 0$ on a weighted continuous function space, but it also suggests the following scaling on the generalization error,

$$\mathbb{E}_{\tilde{\pi}^\delta} \left[\|(b(X) - b_\epsilon(X))\|^2 \right] \leq K_3^2 \mathbb{E}_{\tilde{\pi}^\delta} \left[1 + \|X\|^2 \right] \epsilon^2 = O(\epsilon^2). \tag{26}$$

This implies that whenever the generalization error is scaled as in (26), which will always be satisfied when the condition (25) is valid, the spectral error of the diffusion matrix estimator (with error bound in (23)) is dominated by the Monte-Carlo error that can be controlled by concentration inequalities, instead of the generalization error in learning b.

To deduce the result below, we adopt the notation in Section 2. For $\delta > 0$, let X_n^ϵ and $X_n^{\epsilon,\delta}$ denote the Markov chain sampled from the perturbed dynamics (15) with $X_n^\epsilon = X^\epsilon(n\delta)$ and the discretized Markov chain of (15) generated by EM scheme (7) with step size δ, respectively. Here, $\{X_n^\epsilon\}$ and $\{X_n^{\epsilon,\delta}\}$ can be interpreted as the perturbed Markov chains of X_n and X_n^δ, respectively. To employ the existing theory reviewed in Section 2, recall that Proposition 2.1 in Section 2.1 involves an unperturbed Markov chain $\{u_n\}$ and the corresponding perturbed Markov chain $\{u_n^\epsilon\}$, where $\{u_n\}$ is assumed to be geometrically ergodic. While it is obvious that we are fundamentally interested in the case where $(u_n, u_n^\epsilon) = (X_n, X_n^\epsilon)$, in practice, we are rarely given a realization of $\{X_n\}$. This implies that when the drift and diffusion coefficients are estimated, we will not be able to realize samples of X_n^ϵ, the estimated dynamics, that are independent of time discretization, $\delta > 0$, even if one can theoretically write the error bound between of the statistics of X_n and X_n^ϵ (as we shall see below). What is usually available is the time series of $\{X_n^\delta\}$, obtained e.g., via EM integrator. Given such constraints, we consider also the case where $(u_n, u_n^\epsilon) = (X_n^\delta, X_n^{\epsilon,\delta})$, in which $X_n^{\epsilon,\delta}$ denotes the estimated Markov chain with the drift and diffusion coefficients obtained by learning from the time series $\{X_n^\delta\}$. By Assumption 2.3 and Proposition 2.2, we know that the Markov chain $\{X_n^\delta\}$ is geometrically ergodic and the condition related to the Lyapunov function (2) is valid for both two cases. Thus, it is enough to derive the relation between the γ in (3) and the parameter ϵ, which is given by the following lemma.

Lemma 3.1. Consider the unperturbed ergodic Itô diffusion (9) and the corresponding perturbed dynamics (15) satisfying Assumptions 2.1–2.3 and Assumption 3.1. For the fixed step size δ specified in Theorem 2.1 and all $0 < \epsilon \ll 1$, we have

$$\gamma_X := \sup_{x \in \mathbb{R}^d} \sup_{f \in \mathscr{G}_\ell} \frac{\left| \mathbb{E}^x[f(X_1)] - \mathbb{E}^x[f(X_1^\epsilon)] \right|}{V(x)} \leq K\epsilon,$$

$$\gamma_{X^\delta} := \sup_{x \in \mathbb{R}^d} \sup_{f \in \mathscr{G}_\ell} \frac{\left| \mathbb{E}^x[f(X_1^\delta)] - \mathbb{E}^x[f(X_1^{\epsilon,\delta})] \right|}{V(x)} \leq K^\delta \epsilon,$$

for some constant $K, K^\delta \in (0, +\infty)$ that are independent of ϵ. Here, $\mathscr{G}_\ell$ is defined as in Proposition 2.2.

See Appendix A for the proof. We should point out that in the course of the proof, the linear scaling in (25) is important for balancing the ϵ scaling that is defined as the spectral error of the estimator $\sigma_\epsilon \sigma_\epsilon^\top$ in (24). The proof also suggests that while the scaling of generalization error in (26) is necessary for small spectral error bound in (24) as we pointed out right after (26), it is not a sufficient condition to achieve the bounds in Lemma 3.1. This issue is due to the discrepancy of topologies defined in the perturbation theory and the usual generalization error in machine learning. Specifically, the perturbation theory result used in this paper to achieve the linear dependence result is defined with a stronger topology than that of the generalization error in learning theory.

3.2. One-point statistics

For one-point statistics, we consider an observable $f : \mathbb{R}^d \to \mathbb{R}$ with finite first moment, $\pi(|f|) < \infty$, and derive error bounds for $\left|\mathbb{E}^x[f(X_n^\epsilon)] - \pi(f)\right|$ and $\left|\mathbb{E}^x[f(X_n^{\epsilon,\delta})] - \pi(f)\right|$, where π denotes the invariant measure of the unperturbed dynamics (9). By Lemma 3.1, if the observable $f \in \mathscr{G}_\ell$ (defined in Proposition 2.2), such bounds can be derived immediately from Proposition 2.1 and they are uniform for all $f \in \mathscr{G}_\ell$. In particular, the following proposition summarizes the corresponding results.

Proposition 3.1. *Under the same circumstances as in Lemma 3.1, consider the invariant measure of the unperturbed dynamics (9), π. Let $\{X_n^\epsilon\}$ denote the Markov chain of a generic perturbation of size ϵ. Furthermore, we let $\{X_n^{\epsilon,\delta}\}$ denote the Markov chain generated by the EM integration of (15). Then, we have*

$$\sup_{f \in \mathscr{G}_\ell} \left|\mathbb{E}^x[f(X_n^\epsilon)] - \pi(f)\right| \le R_1\left[\left(\rho_1^n + \frac{1-\rho_1^n}{1-\rho_1}\epsilon\right)V(x)\right], \quad \forall n \ge 0,$$

and

$$\sup_{f \in \mathscr{G}_\ell} \left|\mathbb{E}^x[f(X_n^{\epsilon,\delta})] - \pi(f)\right| \le R_2\left[\left(\rho_2^n + \frac{1-\rho_2^n}{1-\rho_2}\epsilon\right)V(x) + \delta^\nu\pi(V)\right], \quad \forall n \ge 0, \tag{27}$$

for some constants $R_1, R_2 \in (0, +\infty)$, $\rho_1, \rho_2 \in (0, 1)$, and $\nu \in (0, \frac{1}{2})$ (same as in Eq. (12)).

Proof. Both inequalities are results of Eq. (5), which is a corollary of Proposition 2.1. By Lemma 3.1, we know Proposition 2.1 holds for both $(u_n, u_n^\epsilon) = (X_n, X_n^\epsilon)$ and $(u_n, u_n^\epsilon) = (X_n^\delta, X_n^{\epsilon,\delta})$. To obtain the first inequality, we apply Eq. (5) to X_n^ϵ (with π being the invariant measure of X_n).

As we have pointed out in the discussion above, while this inequality holds, in practice, we will not have access to X_n^ϵ. As for the second inequality, we apply Eq. (5) to $X_n^{\epsilon,\delta}$ (with $\pi = \tilde{\pi}^\delta$ being the invariant measure of X_n^δ defined as in Proposition 2.2). The desired bounds can be obtained using the triangular inequality together with Eq. (12) in Proposition 2.2. Namely,

$$\left|\mathbb{E}^x[f(X_n^{\epsilon,\delta})] - \pi(f)\right| \le \left|\mathbb{E}^x[f(X_n^{\epsilon,\delta})] - \tilde{\pi}^\delta(f)\right| + \left|\tilde{\pi}^\delta(f) - \pi(f)\right|$$

$$\le R\left[(1-\rho_2^n)\frac{\gamma_{X^\delta}\kappa}{1-\rho_2} + \rho_2^n V(x)\right] + K\delta^\nu\pi(V)$$

To achieve the bound in (27), we use that fact that $V(x) \in [1, +\infty)$ to replace the term κ in Eq. (5) by $V(x)$, and define $R_2 = \max\{RK^\delta, R, K\}$. $\square$

Essentially, this Proposition suggests that the error in one-point statistics depends linearly on the perturbation size, ϵ, as $n \to \infty$. It is worthwhile to note this result is stronger than a similar conclusion in [48] since their result assumes a uniform consistency, $\sup_{x \in \mathbb{R}^d} \|b_\epsilon(x) - b(x)\| \le \epsilon$, which is stronger than our assumption, a linear error growth in (25). In practice, since the available data for training is subjected to numerical discretization error, the error rate is of order $O(\epsilon, \delta^\nu)$, from some $\nu \in (0, 1/2)$ specified in Proposition 2.2. We should point out that ϵ will be inversely proportional to the discretization error, δ, as we shall see in Sections 4.2.3 and 5.2, where this relation is induced by the noises in the response data (16).

3.3. Two-point statistics

For two-point statistics, we consider observables $A, B : \mathbb{R}^d \to \mathbb{R}$ with finite second moments, that is, $\pi(A^2), \pi(B^2) < \infty$. The two-point statistics of A and B for the unperturbed dynamics (9) are defined as

$$k_{A,B}(t) := \mathbb{E}_\pi\left[A(X(t))B(X(0))\right] = \iint A(x)B(x_0)P_t(x_0, \mathrm{d}x)\pi(\mathrm{d}x_0). \tag{28}$$

In applications, the two-point statistics can arise from Fluctuation–Dissipation Theory and is a route to approximate the statistics of a system driven out of equilibrium, e.g., [49–51]. In [52], we have shown that the two-point statistics in (28) are well-defined for all $t \ge 0$. Formally, under Assumption 2.5, we can define the corresponding two-point statistics for the perturbed dynamics (15), as an approximation of (28). It is given by

$$k_{A,B}^\epsilon(t) := \mathbb{E}_{\pi^\epsilon}\left[A(X^\epsilon(t))B(X^\epsilon(0))\right] = \iint A(x)B(x_0)P_t^\epsilon(x_0, \mathrm{d}x)\pi^\epsilon(\mathrm{d}x_0).$$

Here, P_t^ϵ denotes the transition kernel (10) of the perturbed dynamics. In terms of the Markov chains X_n and X_n^ϵ, the two-point statistics reduce to

$$(k_{A,B})_n := k_{A,B}(n\delta) = \mathbb{E}_\pi[A(X_n)B(X_0)], \quad (k_{A,B}^\epsilon)_n := k_{A,B}^\epsilon(n\delta) = \mathbb{E}_{\pi^\epsilon}[A(X_n^\epsilon)B(X_0^\epsilon)], \quad \forall n \ge 0. \tag{29}$$

The following proposition, as the main result of this section, provides an error bound for the two-point statistics.

Proposition 3.2. *Under the same circumstances as in* Lemma 3.1, *for a fixed step size* $\delta > 0$, *let* Assumptions 2.5–2.8 *hold for the family of Markov operators,*

$$\left(\mathscr{P}_\delta^\epsilon f\right)(x) := \int f(y) P_\delta^\epsilon(x, dy) = \mathbb{E}^x[f(X^\epsilon(\delta))], \quad 0 < \epsilon \ll 1,$$

induced by the perturbed dynamics (15) *with respect to the functions* G, H, U *(see* Theorem 2.2 *for the details) satisfying* $G \geq V$. *We further assume*

$$\pi^\epsilon(V^2) < \infty,$$

where V *is the Lyapunov function of the unperturbed dynamics. Then, for any observables* $A \in \mathscr{G}_\ell \cap C^1_{V, VH/G}$ *and* $B \in C^1_{G/V, H/V}$ *satisfying*

$$\pi(A^2), \ \pi(B^2), \ \pi^\epsilon(A^2), \ \pi^\epsilon(B^2) < \infty,$$

the two-point statistics in (29) *are well-defined, and*

$$\left|(k_{A,B}^\epsilon)_n - (k_{A,B})_n\right| \leq R\pi^\epsilon(B^2)^{\frac{1}{2}}\epsilon + \left|\frac{d}{d\epsilon}\pi^\epsilon(f_n)\right|_{\epsilon=0}\left|\epsilon + O(\epsilon^2), \quad f_n(x) = B(x)\mathbb{E}^x[A(X_n)], \right. \tag{30}$$

for some constants $R \in (0, +\infty)$ *independent of* A *and* B. *Here, the* ϵ-*derivative in* (30) *is well-defined and satisfies Eq.* (14) *with* $f = f_n$ *and* $t = \delta$. *Moreover, if we further assume that* A *is centered with respect to* π ($\pi(A) = 0$), *the error bound in* (30) *satisfies*

$$\left|(k_{A,B}^\epsilon)_n - (k_{A,B})_n\right| \leq R\left[\pi^\epsilon(B^2)^{\frac{1}{2}} + \frac{1}{1-\lambda}\|B\|_{1;G/V,H/V}\left(\lambda^n\|A\|_{1;V,VH/G} + \rho^n\pi(G)\right)\right]\epsilon + O(\epsilon^2), \tag{31}$$

for some constants $R \in (0, +\infty)$, $\lambda \in (0, 1)$ *(the same as those in* Assumption 2.6*), and* $\rho \in (0, 1)$ *(the same as in* Theorem 2.1*) independent of* A *and* B. *The norms in* (31) *are defined by Eq.* (13).

See Appendix B for the proof. In the proof we analyze the two factors that contribute to the error of the two-point statistics: the transition kernel and the invariant measure, which lead to the first and second terms on the right-hand side of the error bound (30), respectively. When the observable A is centered with respect to π, suggested by Eq. (31), the leading order term of the error caused by the invariant measure, goes to zero as $n \to +\infty$.

Proposition 3.2 states the error of the two-point statistics between the two sampling Markov chains $\{X_n\}$ and $\{X_n^\epsilon\}$. Similar to Lemma 3.1 and Proposition 3.1, one can extend the error bounds in Proposition 3.2 corresponding to the two discretized Markov chains: $\{X_n^\delta\}$ and $\{X_n^{\epsilon,\delta}\}$, where the two-point statistics are defined with respect to the discrete transition kernel. For example, the two-point statistics of $\{X_n^\delta\}$ are defined as

$$\left(k_{A,B}^\delta\right)_n := \mathbb{E}_{\tilde{\pi}^\delta}\left[A(X_n^\delta)B(X_0^\delta)\right] = \iint A(x)B(x_0)(P^\delta)^n(x_0, dx)\tilde{\pi}^\delta(dx), \quad \forall n \geq 0,$$

where $A, B \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ and P^δ denotes the transition kernel of $\{X_n^\delta\}$. Here, $(P^\delta)^n$ corresponds to the product of the transition kernel, that is,

$$(P^\delta)^n(x, A) = \int (P^\delta)^{n-1}(x, dy)P^\delta(y, A), \quad \forall A \in \mathscr{B}(\mathbb{R}^d), \quad \forall n \geq 1.$$

Remark 1. Proposition 3.2 is valid under a series of assumptions, which can be classified into the following three categories.

  i. **Assumptions on the unperturbed dynamics:** Assumptions 2.1–2.4 are proposed to ensure the geometric ergodicity of the unperturbed dynamics and its numerical discretization.
 ii. **Assumptions on the perturbed dynamics:** Assumption 3.1 is postulated so that the perturbed dynamic is a "reasonable" approximation of the unperturbed dynamics (Lemma 3.1). Assumptions 2.5–2.8 are proposed so that the long-time linear response theory holds (Theorem 2.2).
iii. **Assumptions on the observables**: We assume the observables A and B satisfy certain regularity and integrability conditions, so that the two-point statistics in (29) are well-defined and the long-time linear response theory is applicable to f_n in (30). The regularity assumptions of A and B are not identical. Specifically, $A \in C^1_{V, VH/G}$ while $B \in C^1_{G/V, H/V}$. In particular, when $G \geq V^2$, we have

$$G/V \geq V, \quad H/V \geq VH/G,$$

which imply that $C^1_{V, VH/G} \subset C^1_{G/V, H/V}$, that is, the regularity assumption of B is weaker than that of A. In practice, it is common that the observable B is not as regular as the observable A. For example, in the Fluctuation–Dissipation theory (FDT), the linear response operator defines two-point statistics of the form in (28) between the observable A and the conjugate variable B [53]. The conjugate variable B, produced by a differential operator (a typical example is $B = \nabla \log(\pi)$ [54]), is often less regular than A.

Among all the assumptions, the assumptions associated with the long-time linear response theory (Assumptions 2.5–2.8) are somewhat abstract and difficult to be directly verified. To provide some insight, we give an example here to show how the long-time linear response theory is applied to a class of Itô diffusions and their perturbations.

Example 3.1. Consider an Itô diffusion (9) satisfying Assumptions 2.1–2.2. If the drift coefficient b in (9) satisfies the dissipative condition [39], that is, there exist constants $a, d \in (0, +\infty)$ such that

$$\langle b(x), x \rangle \leq a - d\|x\|^2, \quad \forall x \in \mathbb{R}^d,$$

then, Assumptions 2.3–2.4 hold with $V(x) = 1 + \|x\|^{2\ell}$ for all $\ell \geq 1$. Moreover, consider the family of Markov operators $\{\mathscr{P}_t^\epsilon\}$ induced by the perturbed dynamics in (15) with coefficients satisfying Assumption 3.1, then Assumptions 2.5–2.8 hold for all ϵ small enough with

$$G(x) = 1 + \alpha e^{\eta \|x\|^2}, \quad H(x) = \beta^{-1} e^{\eta \|x\|^2}, \quad U(x) = G(x) + e^{2\eta \|x\|^2},$$

for some positive constants α, β, and η sufficiently small.

The choice of the Lyapunov function is a result of Lemma 4.2 in [39]. Since the perturbation to the drift coefficients $b_\epsilon - b$ is of linear growth (Assumption 3.1), for ϵ small enough, the drift coefficients $b_\epsilon = b + (b_\epsilon - b)$ of the perturbed dynamics still satisfy the dissipative condition. Under such observations, the choice of functions G, H and U can be found in the proof of Theorem 4.4 in [40]. In [40], the authors considered the case where the drift coefficient b is a linear combination of symmetric multi-linear maps, so that they can apply the long-time linear response theory to the family of Markov operators parameterized by the parameters in the SDEs. In our situation, the family of Markov operators is parameterized by ϵ in the perturbed dynamics. With the key ingredient, the dissipative condition, being preserved, the proof in [40] is still valid. It is worthwhile to mention that, given the dissipative condition, Assumption 2.2(i) can be replaced by the assumption that the dynamics (9) is approximately controllable [39,40]. As for the existence of the invariant measure π^ϵ, one can also consider the stationary Fokker–Planck equation of (15). In our case, since the noise may be degenerate, the existence of π^ϵ can be established by results in [55] with the help of the Lyapunov function.

4. Learning with the kernel-based spectral regression method

In Section 3, we have deduced a linear dependence of the error of the invariant statistics to the error in the estimation of the drift and diffusion coefficients. Our primary interest here (and in Section 5) is to understand under which conditions can the results in Section 3 be achieved when existing popular machine learning methods are used to estimate the coefficients b and $\sigma \sigma^\top$. Here, we will focus on the kernel-based spectral regression method [41], whereas, in the next section, we will focus on the single hidden-layer random neural networks with the ReLU activation functions [30].

In Section 3.1, we have formally proposed a framework for learning ergodic SDEs by solving a regression problem and introduced the concept of the generalization error, which will be further decomposed into two parts: estimation error (error caused by sampling) and approximation error (error due to the choice of hypothesis space). In particular, we have pointed out that the coefficients $(b_\epsilon, \sigma_\epsilon \sigma_\epsilon^\top)$ in the approximated system (15) are the empirical estimates defined through (20) for a given sample. Using the notation in (18), our goal now is to quantify the generalization error defined in (26) induced by b_ϵ, an estimator obtained from training on a set of i.i.d. labeled data, sampled from (X, Y) generated by (16). Through the assumed scaling in (26), this analysis elucidates how the parameter ϵ depends on the training sample size, the parameters in the hypothesis space, noise amplitude parameter, and step size for the time discretization.

To simplify the discussion, we will perform the analysis component-wise. Abusing the notation, we refer $b : \mathbb{R}^d \to \mathbb{R}$ as a generic notation for each component of $b : \mathbb{R}^d \to \mathbb{R}^d$. Correspondingly, we now refer to $y_i := b(x_i) + \eta_i \in \mathbb{R}$ as a generic component of a (d-dimensional vector) sample of Y in (16). Here, η_i denotes a real-valued component of the d-dimensional Gaussian random variable $\mathcal{N}(0, \delta^{-1} \sigma \sigma^\top)$. With this abuse of notation, our training data set is denoted by the labeled data $\{x_i, y_i\}_{i=1}^N$ with $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$.

For completeness, we will review some basic concepts of reproducing kernel Hilbert spaces (RKHS) in Section 4.1. Subsequently, in Section 4.2, we will discuss a data-driven kernel-based spectral regression approach, whose mathematical foundation lies in the theory RKHSs. While controlling the generalization error bound ϵ is a practical interest, as we pointed out in Section 3.1, this is only a necessary and not a sufficient condition for achieving the main results (Propositions 3.1 and 3.2) in this paper. Therefore, it is crucial to understand whether the conditions in Assumption 3.1, the linear growth bound and globally Lipschitz continuity, which are the necessary conditions to the results in Section 3, can be satisfied (see Corollary 4.1 and Proposition 4.2). We close this section with some discussions of the advantages and shortcomings of this estimation approach in Section 4.3.

4.1. A brief review of RKHS

For simplicity, we will consider the class of RKHS of the real valued function spaces on $\mathbb{R}^d$, while the argument can be extended to general locally compact metric spaces [56]. To begin with, recall that a function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a (Mercer) kernel [56,57] if it is continuous, symmetric and positive semidefinite, i.e.,

$$K_x := K(\cdot, x) \in C(\mathbb{R}^d), \quad \forall x \in \mathbb{R}^d,$$

and for any finite set of points $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ the matrix

$$\boldsymbol{K}_N := \frac{1}{N} \big(K(x_i, x_j) \big)_{i,j=1}^N \in \mathbb{R}^{N \times N}, \tag{32}$$

is symmetric positive semidefinite. In the literature, such a matrix (32) is called the *empirical kernel* [41] of K with respect to the sample points $\{x_i\}_{i=1}^N$ and the map $\Phi : x \to K_x$ is called the *feature map* [56,58].

Using the feature map, the RKHS $\mathcal{H}$ associated with the kernel K is defined to be the closure of $\mathrm{span}\{K_x : x \in \mathbb{R}^d\}$ [57] with the inner product given by

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^{M_1} \sum_{j=1}^{M_2} c_i d_j K(x_i, y_j), \quad f = \sum_{i=1}^{M_1} c_i K_{x_i}, \quad g = \sum_{j=1}^{M_2} d_j K_{y_j}.$$

The reproducing property takes the form

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}, \quad \forall x \in \mathbb{R}^d, \tag{33}$$

which implies that $\mathscr{H}$ consists of continuous functions on $\mathbb{R}^d$. As a result of the reproducing property, RKHS has a remarkable property that the $\mathscr{H}$-norm convergence implies pointwise convergence since the evaluation functionals are bounded. Let $\|\cdot\|_{\mathscr{H}}$ denote the norm in the RKHS $\mathscr{H}$. To develop the orthonormal basis of $\mathscr{H}$ for our applications, we propose the following assumption on the kernel.

Assumption 4.1. The kernel K is Hilbert–Schmidt with respect to the nondegenerate probability measure $\tilde{\pi}^\delta$, that is,

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (K(x,y))^2 \mathrm{d}\tilde{\pi}^\delta(x)\mathrm{d}\tilde{\pi}^\delta(y) < +\infty, \tag{34}$$

and $K_x \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ for all $x \in \mathbb{R}^d$. Recall that the available sample points $\{x_i\}_{i=1}^N$ are drawn from $\tilde{\pi}^\delta$.

As a result of Assumption 4.1, the following integral operator

$$L_K : L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \to L^2(\mathbb{R}^d, \tilde{\pi}^\delta), \quad (L_K g)(x) = \int_{\mathbb{R}^d} K(x,s)g(s)\mathrm{d}\tilde{\pi}^\delta(s), \tag{35}$$

is bounded, compact and positive on $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ with countably many positive eigenvalues $\{\lambda_i\}_{i=1}^\infty$ [57]. Such type of integral operators are widely studied in various contexts, e.g., graph Laplacian [59] and diffusion maps [60]. In particular, we have $L_K g \in \mathscr{H}$ for any $g \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ [57]. Thus, for each positive eigenvalue $\lambda_i > 0$, we can take the eigenfunction $u_i \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap \mathscr{H}$ such that

$$L_K u_i = \lambda_i u_i, \quad \langle u_i, u_j \rangle_{\tilde{\pi}^\delta} = \delta_{ij}, \quad \forall i,j \geq 1,$$

where $\langle \cdot, \cdot \rangle_{\tilde{\pi}^\delta}$ denotes the inner product in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. Note that $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ consists of equivalence classes of functions and consequently it is not an RKHS. The fact that for nonzero eigenvalues, one can pick eigenfunctions in $\mathscr{H}$ are critical for our later construction of the estimates.

To see the connection between the eigenfunctions $\{u_i\}$ and the kernel, we introduce the spectral decomposition of L_K [41],

$$L_K = \sum_{i=1}^\infty \lambda_i \langle \cdot, u_j \rangle_{\tilde{\pi}^\delta} u_j.$$

Formally, we can exchange the order of summation and integration in the decomposition above and reach the following representation of the kernel K,

$$K(x,y) = \sum_{i=1}^\infty \lambda_i u_i(x)u_i(y). \tag{36}$$

The relation in (36) is known as the Mercer theorem [56], which is valid even on noncompact domains, e.g., $\mathbb{R}^d$ [57]. As a corollary of the Mercer theorem, $\{\sqrt{\lambda_i}u_i\}_{i=1}^\infty$ form an orthonormal basis of $\mathscr{H}$. In particular, notice that

$$K_x = \sum_{i=1}^\infty \lambda_i u_i(x)u_i,$$

and

$$\langle K_x, K_y \rangle_{\mathscr{H}} = K(x,y) \Rightarrow \sum_{i,j=1}^\infty \sqrt{\lambda_i \lambda_j} u_i(x)u_j(y)\langle \sqrt{\lambda_i}u_i, \sqrt{\lambda_j}u_j \rangle_{\mathscr{H}} = \sum_{i=1}^\infty \lambda_i u_i(x)u_i(y),$$

which leads to the orthogonality properties,

$$\langle \sqrt{\lambda_i}u_i, \sqrt{\lambda_j}u_j \rangle_{\mathscr{H}} = \delta_{ij}, \quad \forall i,j \geq 1.$$

Since $\{u_i\}$ is the orthonormal basis of $(\mathrm{Ker}\, L_K)^\perp$ as a subspace of $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$, we have the following isometric isomorphism

$$L_K^{\frac{1}{2}} : \bar{D}_K \to \mathscr{H}, \tag{37}$$

where $\bar{D}_K = (\mathrm{Ker}\, L_K)^\perp$ is the closure of $D_K = \mathrm{span}\{u_i\}$ in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. Here, $L_K^{\frac{1}{2}}$ is the square-root of L_K satisfying

$$L_K^{\frac{1}{2}} u_i = \sqrt{\lambda_i}u_i, \quad \forall i \geq 1.$$

The isomorphism $L_K^{\frac{1}{2}}$ also reveals the regularity difference between functions in $\bar{D}_K$ and functions in $\mathscr{H}$. We should point out that if all the eigenvalues of L_K are positive, then $L_K^{1/2}$ is injective and $\bar{D}_K$ is dense in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ (e.g., Theorem 4.26 in [56]). This means that any function $f \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ can be approximated with arbitrary precision by a function in the RKHS $\mathscr{H}$, where the convergence is valid in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. As we shall see later, this fact allows one to quantify the approximation error in term of the finite number of basis functions used in the numerical approximation of b.

4.2. Spectral regression with integral operators

In Section 4.1, we have reviewed some basic concepts of RKHS. Given a kernel K that satisfies Assumption 4.1, the corresponding RKHS consists of continuous functions (rather than equivalent classes as in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$) that can be written as linear combinations (possibly infinite) of either images of feature maps $\{K_{x_i}\}$ or orthonormal basis functions $\{\sqrt{\lambda_i}u_i\}$. The latter representation leads to a

statistical learning approach that we shall explore in this section. Throughout the section, $\mathscr{H}$ always denotes the RKHS associated with a kernel K that satisfies Assumption 4.1.

We consider the first M (counting multiplicity) eigenvalues of the integral operator L_K in (35) satisfying

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M > \lambda_{M+1} \geq \cdots.$$

Here, we have assumed that there is a spectral gap between modes M and $M + 1$; that is, $\lambda_M - \lambda_{M+1} > 0$. While this assumption will be used to deduce the estimation error, in practice, one can use the spectral gap to empirically determine the number of modes M in analogous to a dimensionality reduction technique via Principal Component Analysis. Let P_M denote the orthogonal projection from $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ onto the span of the first M eigenfunctions $\{u_i\}_{i=1}^M$, that is,

$$P_M b = \sum_{i=1}^M \langle b, u_i \rangle_{\tilde{\pi}^\delta} u_i, \quad \forall b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta). \tag{38}$$

Recall that $\{u_j\}$ form an orthonormal family in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. For any $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$, the L^2-convergence of $P_M b$ to Pb is clear, where

$$Pb = \sum_{i=1}^\infty \langle b, u_i \rangle_{\tilde{\pi}^\delta} u_i,$$

is the projection onto $\bar{D}_K$. Moreover, if $b \in \mathscr{H}$, such a convergence is valid in $\mathscr{H}$ as well.

Lemma 4.1. *Let K be a kernel that satisfies Assumption 4.1 and $\mathscr{H}$ be the corresponding RKHS. Then $\mathscr{H} \subset L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$, and $\forall b \in \mathscr{H}$, we have $P_M b \to b$ as $M \to \infty$ in $\mathscr{H}$ and*

$$|b(x) - P_M b(x)| \leq \|b - P_M b\|_{\mathscr{H}} K^{\frac{1}{2}}(x, x), \quad \forall x \in \mathbb{R}^d,$$

where $P_M b$ is the projection of b defined in (38).

Proof. The inclusion $\mathscr{H} \subset L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ is a result of Assumption 4.1 (see e.g. [57]). As a result, $\forall b \in \mathscr{H}$ the projection $P_M b$ in (38) is well-defined. We first show that $Pb = b$. By the isometric isomorphism in (37), we have $L_K^{-\frac{1}{2}} b \in \bar{D}_K$, that is,

$$L_K^{-\frac{1}{2}} b = P\left(L_K^{-\frac{1}{2}} b\right) = \sum_{i=1}^\infty \langle L_K^{-\frac{1}{2}} b, u_i \rangle_{\tilde{\pi}^\delta} u_i = \sum_{i=1}^\infty \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}}{\sqrt{\lambda_i}} u_i,$$

which leads to

$$b = L_K^{\frac{1}{2}} L_K^{-\frac{1}{2}} b = \sum_{i=1}^\infty \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}}{\sqrt{\lambda_i}} L_K^{\frac{1}{2}} u_i = \sum_{i=1}^\infty \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}}{\sqrt{\lambda_i}} \sqrt{\lambda_i} u_i = Pb.$$

Since $\{\sqrt{\lambda_i} u_i\}_{i=1}^\infty$ is an orthonormal basis of $\mathscr{H}$, we have

$$\|b\|_{\mathscr{H}}^2 = \sum_{i=1}^\infty \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}^2}{\lambda_i} < \infty.$$

In particular,

$$b - P_M b = \sum_{i=M+1}^\infty \langle b, u_i \rangle_{\tilde{\pi}^\delta} u_i = \sum_{i=M+1}^\infty \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}}{\sqrt{\lambda_i}} \sqrt{\lambda_i} u_i,$$

which shows that,

$$\|b - P_M b\|_{\mathscr{H}}^2 = \sum_{i=M+1}^\infty \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}^2}{\lambda_i} \to 0,$$

as $M \to \infty$.

By the reproducing property (33), we have $\|K_x\|_{\mathscr{H}}^2 = \langle K_x, K_x \rangle_{\mathscr{H}} = K(x, x)$ and

$$|b(x) - P_M b(x)| = |\langle b - P_M b, K_x \rangle_{\mathscr{H}}| \leq \|b - P_M b\|_{\mathscr{H}} \|K_x\|_{\mathscr{H}} = \|b - P_M b\|_{\mathscr{H}} K^{\frac{1}{2}}(x, x), \quad \forall x \in \mathbb{R}^d. \quad \square$$

Lemma 4.1 provides a pointwise error bound for the projection $P_M b$ in (38) given $b \in \mathscr{H}$, which is closely related to the error condition (25) in Assumption 3.1. In particular, the following corollary clarifies this relation and provides sufficient conditions for Assumption 4.1.

Corollary 4.1. *Let K be a kernel of linear growth bound, that is,*

$$K(x, x) \leq C(1 + \|x\|^2), \quad \forall x \in \mathbb{R}^d, \tag{39}$$

for some constant $C \in (0, +\infty)$. Then, K satisfies Assumption 4.1 with respect to the probability measure $\tilde{\pi}^\delta$. Moreover, let $\mathscr{H}$ be the corresponding RKHS, we have $\forall b \in \mathscr{H}$

$$|b(x) - P_M b(x)|^2 \leq C(1 + \|x\|^2) \|b - P_M b\|_{\mathscr{H}}^2, \quad \forall x \in \mathbb{R}^d, \tag{40}$$

which is analogous to Eq. (25) in Assumption 3.1.

Proof. To see the kernel K satisfies the Hilbert–Schmidt condition in (34), notice that

$$(K(x, y))^2 = \langle K_x, K_y \rangle_{\mathscr{H}}^2 \leq \|K_x\|_{\mathscr{H}}^2 \|K_y\|_{\mathscr{H}}^2 = K(x, x)K(y, y),$$

such that,

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (K(x, y))^2 \mathrm{d}\tilde{\pi}^\delta(x)\mathrm{d}\tilde{\pi}^\delta(y) \leq \left(\int_{\mathbb{R}^d} K(x, x)\mathrm{d}\tilde{\pi}^\delta(x) \right)^2 \leq C^2 \left(\int_{\mathbb{R}^d} (1 + \|x\|^2)\mathrm{d}\tilde{\pi}^\delta(x) \right)^2 < +\infty,$$

where we have used the linear growth bound of the kernel. Here, $\mathbb{E}_{\tilde{\pi}^\delta}[\|x\|^2] < \infty$ due to the Lyapunov property of the process X_n^δ (see Proposition 2.2 for the details).

On the other hand, $\forall x \in \mathbb{R}^d$, we have

$$\|K_x\|_{\tilde{\pi}^\delta}^2 = \int_{\mathbb{R}^d} (K(x, y))^2 \mathrm{d}\tilde{\pi}^\delta(y) \leq K(x, x) \int_{\mathbb{R}^d} K(y, y)\mathrm{d}\tilde{\pi}^\delta(y) \leq C^2(1 + \|x\|^2) \int_{\mathbb{R}^d} (1 + \|y\|^2)\mathrm{d}\tilde{\pi}^\delta(y) < +\infty,$$

that is, $K_x \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. Thus, K satisfies Assumption 4.1, and the bound in (40) is a result of Lemma 4.1. □

Example 4.1 (*Polynomial kernels*). The polynomial kernel of degree-1 [56],

$$P_1(x, y) = x^\top y + d, \quad d > 0, \tag{41}$$

naturally satisfies the condition in Corollary 4.1. For functions in the RKHS associated with the polynomial kernel $K = P_1$ (41), we can choose the Lipschitz constant to be proportional to the RKHS norm. In particular, by the reproducing property, we have $\forall b \in \mathscr{H}$,

$$|b(x) - b(y)| = |\langle b, K_x - K_y \rangle_{\mathscr{H}}| \leq \|b\|_{\mathscr{H}} \|K_x - K_y\|_{\mathscr{H}},$$

where

$$\|K_x - K_y\|_{\mathscr{H}} = \langle K_x - K_y, K_x - K_y \rangle_{\mathscr{H}}^{\frac{1}{2}} = (P_1(x, x) + P_1(y, y) - 2P_1(x, y))^{\frac{1}{2}} = \|x - y\|.$$

Thus, in this case, we have

$$|b(x) - b(y)| \leq \|b\|_{\mathscr{H}} \|x - y\|,$$

that is, all $b \in \mathscr{H}$ are globally Lipschitz with Lipschitz constants $\|b\|_{\mathscr{H}}$.

In practice, we can employ other type of kernels with linear growth tail. For example, we can implement a mixed-type kernel, by adding the degree-1 polynomial kernel as in Example 4.1 with kernels that have tail behavior slower than linear growth (mixed-type kernels have been commonly used, see [61]).

4.2.1. Nyström interpolation

In applications, the projection $P_M b$ in (38) is not a practical estimate of b since the eigenfunctions $\{u_j\}$ are unknown in general. Another issue is that assuming $b \in \mathscr{H}$ is too optimistic. Of course, if we are given the information such as the linear growth condition in Assumption 2.1, then we should choose a kernel that also satisfies the condition (as in Corollary 4.1), such as the polynomial kernel in Example 4.1. Without a priori information, the best we can hope is that $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$. To resolve these issues, we first need to come up with empirical estimates of the projection P_M. Then, we should study the properties, including the convergence and the Lipschitz continuity, of the resulting estimates under a mild assumption that $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$.

We are going to construct the estimates of the projection P_M based on the eigenvalues and eigenvectors of the empirical kernel $\boldsymbol{K}_N$ (32) given by the i.i.d. sample points $\{x_i\}$ according to $\tilde{\pi}^\delta$. Such a spectral projection method (onto the data-driven basis constructed by eigen-spaces of the kernel integral operator) has been advocated and widely used in many applications. In the context of learning dynamical systems, see [15–17,62] and the references therein. Assume $r_N = \mathrm{rank}(\boldsymbol{K}_N) \geq M$, and denote $\{\hat{\lambda}_j\}_{j=1}^{r_N}$ as the set of all nonzero eigenvalues (in descending order, counting multiplicity) of $\boldsymbol{K}_N$ with the corresponding normalized eigenvectors $\{\hat{u}_j\} \subset \mathbb{R}^N$, which form an orthonormal family under the inner product of $\langle \cdot, \cdot \rangle_{\tilde{\pi}_N^\delta} := \frac{1}{N} \langle \cdot, \cdot \rangle$. Here $\tilde{\pi}_N^\delta$ denotes the delta measure corresponding to discrete samples $\{x_i\}_{i=1}^N$, that is,

$$\tilde{\pi}_N^\delta = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}.$$

The main difficulty in relating L_K and $\boldsymbol{K}_N$ is that they operate on different spaces. To resolve the issue, we follow [41] and introduce $\hat{v}_j \in \mathscr{H}$ as

$$\hat{v}_j = \frac{1}{N\sqrt{\hat{\lambda}_j}} \sum_{i=1}^N (\hat{u}_j)_i K_{x_i}, \tag{42}$$

where $(\hat{u}_j)_i$ denotes the ith component of the eigenvector $\hat{u}_j$. Here, the set $\{\hat{v}_j\}$ forms an orthonormal family in $\mathscr{H}$, satisfying

$$(\hat{u}_j)_i = \frac{1}{\sqrt{\hat{\lambda}_j}} \hat{v}_j(x_i), \quad i = 1, 2, \ldots, N,$$

that is, $\hat{v}_j / \sqrt{\hat{\lambda}_j}$ can be interpreted as the (Nyström) interpolation of the vector $\hat{u}_j$ in $\mathscr{H}$. To verify the orthogonality, we have

$$\langle \hat{v}_k, \hat{v}_{k'} \rangle_{\mathscr{H}} = \frac{1}{N^2 \sqrt{\hat{\lambda}_k \hat{\lambda}_{k'}}} \sum_{i,j=1}^{N} (\hat{u}_k)_i (\hat{u}_{k'})_j K(x_i, x_j) = \frac{1}{N \sqrt{\hat{\lambda}_k \hat{\lambda}_{k'}}} \hat{u}_k^\top \boldsymbol{K}_N \hat{u}_{k'} = \delta_{kk'}.$$

To this end, we define the following empirical approximation of P_M,

$$\hat{P}_M b = \sum_{i=1}^{M} \langle b, \hat{v}_i \rangle_{\mathscr{H}} \hat{v}_i, \tag{43}$$

where

$$\langle b, \hat{v}_j \rangle_{\mathscr{H}} = \frac{1}{N \sqrt{\hat{\lambda}_j}} \sum_{i=1}^{N} (\hat{u}_j)_i \langle b, K_{x_i} \rangle_{\mathscr{H}} = \frac{1}{\sqrt{\hat{\lambda}_j}} \langle R_N b, \hat{u}_j \rangle_{\tilde{\pi}_N^\delta}, \quad R_N b := (b(x_1), b(x_2), \ldots, b(x_N))^\top.$$

Here, $R_N : \mathscr{H} \to \mathbb{R}^N$ is called the sampling operator [41,63] associated with the discrete set $\{x_i\}_{i=1}^{N}$. Note that the inner product $\langle R_N b, \hat{u}_j \rangle_{\tilde{\pi}_N^\delta}$ is well-defined for a general function $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$ (continuity is necessary so that the sampling operator R_N is well-defined). Thus, we can extend the definition of $\hat{P}_M$ in (43) to

$$\hat{P}_M b = \sum_{i=1}^{M} \langle R_N b, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta} \frac{\hat{v}_i}{\sqrt{\hat{\lambda}_i}}. \tag{44}$$

Here, the projection $\hat{P}_M$ maps functions in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$ to a finite-dimensional space $\hat{\mathscr{H}}_M := \text{span}\{\hat{v}_i\}_{i=1}^{M} \subset \mathscr{H}$. We will use Eq. (44), instead of Eq. (43), as the definition of the empirical estimates $\hat{P}_M b$ with respect to the projection $P_M b$.

4.2.2. Estimation error

Compared with $\mathscr{H}$, the hypothesis space $\hat{\mathscr{H}}_M$ is of dimension M (the order of the estimates) regardless of the sample size. Note that $\hat{\mathscr{H}}_M$ still depends on the samples due to the choice of the basis functions $\{\hat{v}_j\}$ in (42). We should interpret $\hat{\mathscr{H}}_M$ as the empirical approximation of the underlying hypothesis space $\mathscr{H}_M := \text{span}\{\sqrt{\lambda_i} u_i\}_{i=1}^{M}$, which is independent of the sample but unknown.

As for our estimation, recall that the training data is given by $\{x_i, y_i\}_{i=1}^{N}$ with $y_i = b(x_i) + \eta_i$, where $\{\eta_i\}$ are i.i.d. mean zero Gaussian noise of finite variance. Denoting the random variables $\eta_i \sim E : \Omega \to \mathbb{R}$ with Gaussian probability distribution, $\mathcal{N}(0, R)$, where $R \leq \delta^{-1} \max_{1 \leq i \leq d} \{(\sigma \sigma^\top)_{ii}\}$, and $(\eta_1, \eta_2, \ldots, \eta_N) \sim E_N : \Omega \to \mathbb{R}^N$, our empirical estimate corresponds to,

$$\hat{P}_M(b + E) = \hat{P}_M b + \hat{P}_M E, \quad \hat{P}_M E(x) = \sum_{i=1}^{M} \langle E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta} \frac{\hat{v}_i(x)}{\sqrt{\hat{\lambda}_i}}. \tag{45}$$

Thus, the estimation error $P_M b - \hat{P}_M(b + E)$ consists of the error introduced by the empirical projection, $P_M b - \hat{P}_M b$ and the error induced by the noise in the observation, $\hat{P}_M E$. Notice that,

$$\|\hat{P}_M E\|_{\tilde{\pi}^\delta}^2 = \sum_{i=1}^{M} \langle E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}^2 \leq \|E_N\|_{\tilde{\pi}_N^\delta}^2 = \frac{1}{N} \sum_{i=1}^{N} \eta_i^2,$$

follows a χ^2-distribution of degree N. By the following concentration inequality (e.g., Example 2.11 in [64]),

$$\mathbb{P}\left(\left| \frac{1}{NR} \sum_{i=1}^{N} \eta_i^2 - 1 \right| \geq t \right) \leq 2 e^{-\frac{Nt^2}{8}}, \quad \forall t \in (0, 1),$$

we conclude that for any $\tau > 0$ and N large enough,

$$\|\hat{P}_M E\|_{\tilde{\pi}^\delta}^2 \leq \left(1 + \sqrt{\frac{8\tau}{N}} \right) R, \quad R \leq \delta^{-1} \text{Tr}[\sigma \sigma^\top], \tag{46}$$

with probability greater than $1 - 2 e^{-\tau}$. The norm $\| \cdot \|_{\tilde{\pi}_N^\delta}$ is defined with respect to the inner product $\langle \cdot, \cdot, \rangle_{\tilde{\pi}_N^\delta}$.

For bounded kernels, the following proposition characterizes the convergence of $\|P_M b - \hat{P}_M(b + E)\|_{\tilde{\pi}^\delta}$ for any $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$.

Proposition 4.1. Let $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$ and given the training data $\{x_i, y_i\}_{i=1}^{N}$ as we previously proposed. Suppose that K is a bounded kernel satisfying

$$k_{max} := \sup_{x \in \mathbb{R}^d} K^{\frac{1}{2}}(x, x) < \infty,$$

and let $\hat{P}_M(b + E)$ in (45) be the empirical estimate to the projection $P_M b$ in (38). For any $\tau > 0$, we have

$$\left\| P_M b - \hat{P}_M(b + E) \right\|_{\tilde{\pi}^\delta}^2 \leq \frac{32 k_{max}^2 \tau_1 \|b\|_{\tilde{\pi}_N^\delta}^2}{(\lambda_{M+1} - \lambda_M)^2 \hat{\lambda}_{r_N} N} + \left(1 + \sqrt{\frac{8\tau_2}{N}} \right) \delta^{-1} \text{Tr}[\sigma \sigma^\top], \tag{47}$$

with probability greater than $1 - 2e^{-\tau}$, where $\tau = \min\{\tau_1, \tau_2\}$, given the number N of samples satisfies

$$N > \frac{128 k_{max}^2 \tau_1}{(\lambda_M - \lambda_{M+1})^2}.$$

Here, $\{\lambda_i\}_{i=1}^\infty$ and $\{\hat{\lambda}_i\}_{i=1}^{r_N}$ are the positive eigenvalues of the integral operator L_K and the empirical kernel $\boldsymbol{K}_N$, respectively, with $r_N = \text{rank}(\boldsymbol{K}_N)$.

Proof. We first observe that,

$$\left\| P_M b - \hat{P}_M(b + E) \right\|_{\tilde{\pi}^\delta}^2 \leq \left\| P_M b - \hat{P}_M b \right\|_{\tilde{\pi}^\delta}^2 + \| P_M E \|_{\tilde{\pi}^\delta}^2$$

$$= \left\| P_M b - P_M \hat{P}_M b \right\|_{\tilde{\pi}^\delta}^2 + \left\| (I - P_M)\hat{P}_M b \right\|_{\tilde{\pi}^\delta}^2 + \| P_M E \|_{\tilde{\pi}^\delta}^2,$$

where we have used the Pythagorean theorem to deduce the equality above.

For the first two terms, we have,

$$\left\| P_M b - P_M \hat{P}_M b \right\|_{\tilde{\pi}^\delta}^2 = \left\| P_M (I - \hat{P}_M) b \right\|_{\tilde{\pi}^\delta}^2 \leq \left(\sum_{i=M+1}^{r_N} \frac{\langle R_N b, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}^2}{\hat{\lambda}_i} \right) \left(\sum_{i=M+1}^{r_N} \| P_M \hat{v}_i \|_{\tilde{\pi}^\delta}^2 \right)$$

and

$$\left\| (I - P_M)\hat{P}_M b \right\|_{\tilde{\pi}^\delta}^2 \leq \left(\sum_{i=1}^{M} \frac{\langle R_N b, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}^2}{\hat{\lambda}_i} \right) \left(\sum_{i=1}^{M} \left\| (I - P_M)\hat{v}_i \right\|_{\tilde{\pi}^\delta}^2 \right), \quad i = 1, 2, \ldots, r_N,$$

by Eq. (44). Notice that

$$\left(\sum_{i=1}^{r_N} \frac{\langle R_N b, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}^2}{\hat{\lambda}_i} \right) \leq \frac{1}{\hat{\lambda}_{r_N}} \sum_{i=1}^{r_N} \langle R_N b, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}^2 = \frac{1}{\hat{\lambda}_{r_N}} \| b \|_{\tilde{\pi}_N^\delta}^2.$$

Finally, by Theorem 12 in [41], we have

$$\left\| P_M b - \hat{P}_M b \right\|_{\tilde{\pi}^\delta}^2 \leq \frac{\| b \|_{\tilde{\pi}_N^\delta}^2}{\hat{\lambda}_{r_N}} \left(\sum_{i=1}^{M} \| (I - P_M)\hat{v}_i \|_{\tilde{\pi}^\delta}^2 + \sum_{i=M+1}^{r_N} \| P_M \hat{v}_i \|_{\tilde{\pi}^\delta}^2 \right) \leq \frac{32 k_{max}^2 \tau_1 \| b \|_{\tilde{\pi}_N^\delta}^2}{(\lambda_{M+1} - \lambda_M)^2 \hat{\lambda}_{r_N} N}$$

with probability greater than $1 - 2e^{-\tau_1}$. Together with Eq. (46), the proof is completed. $\square$

In the error bound above, we rely on Theorem 12 in [41] that requires the boundedness of the kernel to apply Hoeffding's inequality. For general kernels satisfying Assumption 4.1 (not necessarily bounded), one can develop similar but much weaker probability bounds via Chebyshev's inequality.

4.2.3. Generalization error

In Section 4.2.2, we have studied the estimation error of the empirical estimate (45) under a model with additive i.i.d. noise. In this section, we will study the approximation error and comment on the generalization error in learning the drift coefficients.

For $b \in L^2(\mathbb{R}^d, \tilde{\pi}^\delta) \cap C(\mathbb{R}^d)$, we introduce the following decomposition

$$b - \hat{P}_M(b + E) = \underbrace{(b - Pb) + (Pb - P_M b)}_{\text{approximation error}} + \underbrace{(P_M b - \hat{P}_M b) + \hat{P}_M E}_{\text{estimation error}}, \tag{48}$$

which is commonly defined in learning theory. The approximation error $\| b - P_M b \|_{\tilde{\pi}^\delta}^2$ satisfies

$$\| b - P_M b \|_{\tilde{\pi}^\delta}^2 = \| b - Pb \|_{\tilde{\pi}^\delta}^2 + \| Pb - P_M b \|_{\tilde{\pi}^\delta}^2, \tag{49}$$

since $(I - P)b \in \ker(L_K)$ and $(P - P_M)b \in \ker(L_K)^\perp$. The term $\| b - Pb \|_{\tilde{\pi}^\delta}^2$ on the right-hand side, independent of M, corresponds to the component of the approximation error induced by the choice of hypothesis space $\mathscr{H}$. This bias is intrinsic in the sense that it only depends on the choice of the kernel. In particular, $\| b - Pb \|_{\tilde{\pi}^\delta} = 0$ if and only if $b \in \ker(L_K)^\perp$. A sufficient condition would be that the integral operator L_K in (35) has only positive eigenvalues. In such a case, $\mathscr{H}$ is dense in $\ker(L_K)^\perp$ with respect to the topology induced by the norm $\| \cdot \|_{\tilde{\pi}^\delta}$. In general, the property of RKHS being dense in a certain function space corresponds to the *universality* of the RKHS [65].

The term $\| Pb - P_M b \|_{\tilde{\pi}^\delta}^2$ describes the approximation error induced by truncation, that is, using $\mathscr{H}_M$, instead of $\mathscr{H}$, as the underlying hypothesis space, which vanishes as the order $M \to \infty$. In particular, we have

$$\| Pb - P_M b \|_{\tilde{\pi}^\delta}^2 = \sum_{i=M+1}^{\infty} \langle b, u_i \rangle_{\tilde{\pi}^\delta}^2,$$

and the decay rate of $\| Pb - P_M b \|_{\tilde{\pi}^\delta}^2$, without further assumption on b, is hard to identify. In our application, the drift coefficients b in (9) is of linear growth bound according to Assumption 2.1. On the other hand, the RKHS associated with a kernel of linear growth bounds (see Corollary 4.1 for the details) consists of functions of linear growth bounds. Thus, it is reasonable to propose the following assumption on b.

Assumption 4.2. Let b be a component of the drift coefficients in (9) satisfying Assumption 2.1. Assume that there exists a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ of linear growth bound (39) and the corresponding RKHS $\mathscr{H}$ such that:

    i. (Decay rate of the eigenvalue) the positive eigenvalues $\{\lambda_i\}$ of the integral operator L_K (35) follow the following decay rate [58],

$$\alpha i^{-r} \leq \lambda_i \leq \beta i^{-r}, \quad \alpha, \beta, r > 0, \quad i = 1, 2, \dots. \tag{50}$$

    ii. (Existence of the target function) the projection sequence $\{P_M b\} \subset \mathscr{H}$ converges to a function $b_{\mathscr{H}}$ in $\mathscr{H}$.

The decay rate assumption in (50) is related to the *effective dimension* [58] of the RKHS $\mathscr{H}$ with respect to the space $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. Recall that by Corollary 4.1, all kernels of linear growth bound satisfy Assumption 4.1, which leads to the inclusion $\mathscr{H} \subset L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. Since the convergence in $\mathscr{H}$ implies the convergence in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$, we have

$$\|Pb - b_{\mathscr{H}}\|_{\tilde{\pi}^\delta} \leq \|Pb - P_M b\|_{\tilde{\pi}^\delta} + \|P_M b - b_{\mathscr{H}}\|_{\tilde{\pi}^\delta} \to 0,$$

as $M \to \infty$, that is, $Pb = b_{\mathscr{H}}$ in $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$. The function $b_{\mathscr{H}}$ is often called the *target function* of b with respect to the hypothesis space $\mathscr{H}$ [44]. With the convergence $P_M b \to b_{\mathscr{H}}$ in $\mathscr{H}$, we have

$$\|b_{\mathscr{H}}\|_{\mathscr{H}}^2 = \lim_{M \to +\infty} \|P_M b\|_{\mathscr{H}}^2 = \sum_{i=1}^{\infty} \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}^2}{\lambda_i} < \infty.$$

As a result, we have

$$\|Pb - P_M b\|_{\tilde{\pi}^\delta}^2 = \sum_{i=M+1}^{\infty} \langle b, u_i \rangle^2 \leq \lambda_{M+1} \sum_{i=M+1}^{\infty} \frac{\langle b, u_i \rangle_{\tilde{\pi}^\delta}^2}{\lambda_i} \leq \lambda_{M+1} \|b_{\mathscr{H}}\|_{\mathscr{H}}^2 = O(M^{-r}), \tag{51}$$

where we have used the decay rate assumption of the eigenvalues (50).

To conclude, in our notation, the estimator $\hat{P}_M(b + E)$ is a component of b_ϵ in (20). Particularly, taking $L^2(\mathbb{R}^d, \tilde{\pi}^\delta)$ on (48) and using (47) and (51), the generalization error is given by,

$$\mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_\epsilon\|^2\right] = O(M^{-r}) + O\left(\frac{1}{(\lambda_{M+1} - \lambda_M)} \hat{\lambda}_{r_N}^{-1} N^{-1}\right) + \delta^{-1} \operatorname{Tr}[\sigma \sigma^\top], \tag{52}$$

as $M, N \to \infty$, where we have assumed that the error rate is uniform component-wise and L_K has only positive eigenvalues such that $\|b - Pb\|_{\tilde{\pi}^\delta} = 0$. Recall that the first term in (52) is the approximation error and the last two terms are estimation errors, respectively, which were derived under various assumptions reported throughout the previous and the current subsections. Importantly, this error bound is valid only for bounded kernels by the assumption in Proposition 4.1. For (unbounded) kernels, e.g., kernels of linear growth, one needs to replace the second error term in (52) with another appropriate rate.

4.2.4. Lipschitz continuity

Now, we check the Lipschitz continuity of the estimator $\hat{P}_M(b + E)$, which is one of the fundamental assumptions (Assumption 3.1) for the statistical error bounds in Propositions 3.1 and 3.2.

Proposition 4.2. Let K be a kernel satisfying Assumption 4.1 (not necessarily bounded). We further assume $K \in C^1(\mathbb{R}^d \times \mathbb{R}^d)$ such that the following function

$$L(x) := \sup_{z \in \mathbb{R}^d} \|\nabla_z K_x(z)\| < \infty, \quad \forall x \in \mathbb{R}^d, \tag{53}$$

is well-defined. Then, given the data set $\{x_i, y_i\}_{i=1}^N$ as in Proposition 4.1, the order-M spectral regression estimates $\hat{P}_M(b + E)$ in (45) satisfies

$$\sup_{x, x' \in \mathbb{R}^d, x \neq x'} \frac{\left|\hat{P}_M(b + E)(x) - \hat{P}_M(b + E)(x')\right|}{\|x - x'\|} \leq \left(\sum_{i=1}^{M} \hat{\lambda}_i^{-2}\right)^{\frac{1}{2}} \left(\|b\|_{\tilde{\pi}_N^\delta} + \|E_N\|_{\tilde{\pi}_N^\delta}\right) \|L\|_{\tilde{\pi}_N^\delta}. \tag{54}$$

Proof. We rewrite the order-M estimates in (44) as

$$\hat{P}_M(b + E)(x) = \frac{1}{N} \sum_{i=1}^{M} \langle R_N b + E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta} \frac{1}{\hat{\lambda}_i} \sum_{j=1}^{N} (\hat{u}_i)_j K(x, x_j) = \sum_{i=1}^{M} \frac{1}{\hat{\lambda}_i} \langle R_N b + E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta} \langle R_N K_x, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}.$$

Thus, for $x, x' \in \mathbb{R}^d$, we have

$$\left|\hat{P}_M(b + E)(x) - \hat{P}_M(b + E)(x')\right| \leq \sum_{i=1}^{M} \left|\frac{1}{\hat{\lambda}_i} \langle R_N b + E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta} \langle R_N K_x - R_N K_{x'}, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}\right|$$

$$\leq \sum_{i=1}^{M} \left|\frac{1}{\hat{\lambda}_i} \langle R_N b + E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}\right| \|R_N K_x - R_N K_{x'}\|_{\tilde{\pi}_N^\delta}.$$

Notice that

$$\sum_{i=1}^{M} \left|\frac{1}{\hat{\lambda}_i} \langle R_N b + E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}\right| \leq \left(\sum_{i=1}^{M} \hat{\lambda}_i^{-2}\right)^{\frac{1}{2}} \left(\sum_{i=1}^{M} \langle R_N b + E_N, \hat{u}_i \rangle_{\tilde{\pi}_N^\delta}^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{M} \hat{\lambda}_i^{-2}\right)^{\frac{1}{2}} \left(\|b\|_{\tilde{\pi}_N^\delta} + \|E_N\|_{\tilde{\pi}_N^\delta}\right).$$

As a result, we have,

$$\left| \hat{P}_M(b+E)(x) - \hat{P}_M(b+E)(x') \right| \leq \left(\sum_{i=1}^{M} \hat{\lambda}_i^{-2} \right)^{\frac{1}{2}} \left(\|b\|_{\tilde{\pi}_N^{\delta}} + \|E_N\|_{\tilde{\pi}_N^{\delta}} \right) \|R_N K_x - R_N K_{x'}\|_{\tilde{\pi}_N^{\delta}}.$$

By the definition of the function $L(x)$, we have

$$|K_x(x_i) - K_{x'}(x_i)| = \left| K_{x_i}(x) - K_{x_i}(x') \right| \leq L(x_i) \|x - x'\|, \quad i = 1, 2, \ldots, N, \quad \forall x, x' \in \mathbb{R}^d.$$

Thus,

$$\sup_{x, x' \in \mathbb{R}^d, x \neq x'} \frac{\left| \hat{P}_M(b+E)(x) - \hat{P}_M(b+E)(x') \right|}{\|x - x'\|} \leq \left(\sum_{i=1}^{N} \hat{\lambda}_i^{-2} \right)^{\frac{1}{2}} \left(\|b\|_{\tilde{\pi}_N^{\delta}} + \|E_N\|_{\tilde{\pi}_N^{\delta}} \right) \left(\frac{1}{N} \sum_{i=1}^{N} L(x_i)^2 \right)^{\frac{1}{2}}$$

$$= \left(\sum_{i=1}^{M} \hat{\lambda}_i^{-2} \right)^{\frac{1}{2}} \left(\|b\|_{\tilde{\pi}_N^{\delta}} + \|E_N\|_{\tilde{\pi}_N^{\delta}} \right) \|L\|_{\tilde{\pi}_N^{\delta}}. \quad \square$$

Here, the function L can be defined for unbounded kernels. For example, for the polynomial kernel $P_1(x, y) = x^\top y + d$, the corresponding $L(x) = \|x\|$. If the kernel K is a radial basis function (RBF) kernel, e.g., Gaussian kernels, the function L in (53) will reduce to a constant function.

Notice that when $N \to \infty$, the upper bound in (54) stays bounded and only depends on the order of the estimates. In particular, when the kernel is bounded, we have the convergence of the eigenvalue as the sample increases, that is, $\hat{\lambda}_i \to \lambda_i$ in high probability as $n \to \infty$ [41]. Under the decay rate assumption in (50), the sum in (54) satisfies (in high probability),

$$\left(\sum_{i=1}^{M} \hat{\lambda}_i^{-2} \right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{M} \alpha^{-1} i^{2r} \right)^{\frac{1}{2}} = O(M^{r+\frac{1}{2}}),$$

as $N \to \infty$. Thus, the Lipschitz constants of the estimates $\hat{P}_M(b+E)$ in (45) is at most of polynomial growth rate with respect to the order of the estimates (in high probability) under the infinite sample assumption.

4.3. Remarks on the spectral regression approach

The spectral regression approach has several advantages. First, one can impose the characteristics of the functions to be estimated in the kernel (such as those in Corollary 4.1). Even when the unknown function to be estimated is unbounded, thanks to the integral operator L_K being compact (as an operator from $L^2(\mathbb{R}^d, \tilde{\pi}^{\delta})$ to itself) with range in the RKHS $\mathscr{H}$, we are allowed to construct the leading M eigenfunctions in $(\mathrm{Ker}(L_K))^{\perp}$ to characterize the RKHS $\mathscr{H}_M \subset \mathscr{H}$ associated with the kernel. One important issue in practice is to identify a kernel such that L_K is strictly positive such that $(\mathrm{Ker}(L_K))^{\perp} = L^2(\mathbb{R}^d, \tilde{\pi}^{\delta})$, which makes the implicit error $\|b - Pb\|_{\tilde{\pi}^{\delta}}^2 = 0$, and hence, does not contribute to the approximation error in (49). The difficulty in specifying such a kernel is attributed to the unknown sampling distribution, $\tilde{\pi}^{\delta}$.

The Nyström interpolation is a convenient tool for associating the eigenvectors of the empirical kernel $\boldsymbol{K}_N$ to the eigenfunctions in $\mathscr{H}$, which leads to the desirable projection that defines our estimates. Notice that each approximated eigenfunction in (42) is still a linear combination of $\{K_{x_i}\}_{i=1}^n$. One advantage of spectral decomposition is that it allows one to separate the effect of finite sample size and the dimension of the hypothesis space, even when the hypothesis space is empirically constructed by interpolating the eigenvectors that depend on the data size. This is in contrast to the general kernel ridge-regression approach [58] with hypothesis spaces that cannot be classified in terms of the dimension. Particularly, when the kernel is radial-type function, there is a lack of ordering in the corresponding set of features $\{K_x(\cdot), \forall x \in \mathbb{R}^d\}$, which is empirically estimated by $\{K_{x_i}(\cdot)\}_{i=1}^N$. By controlling the dimension of the hypothesis space (i.e., fixing the number of basis functions used in the representation), we can easily deduce the Lipschitz continuity as shown in Proposition 4.2.

One practical limitation with the projection-based method is the high computational cost in solving the eigenvalue problem associated with the empirical kernel $\mathbf{K}_N$ for large N. While it is desirable to have a small number of basis functions, M, to remedy this issue, it remains an open question which kernels can induce an RKHS space that can effectively represent the target function with a small number of basis functions. our experience indicates that a careful choice of kernels that also account for the information from the labeled data $\{y_i\}$ in addition to just the covariate data, $\{x_i\}$, is an important direction to pursue in the future study.

5. Learning with ReLU random neural networks

In Section 4, we have discussed the kernel-based spectral regression method in learning the drift coefficients and visited various issues, including the consistency, the generalization error, and the Lipschitz continuity. In particular, the hypothesis space is the span of a finite number of eigenfunctions determined by the Nyström interpolation. The orthogonality of the basis functions provides explicit expressions for the coefficients in minimizing the empirical risk. In this section, we will consider the *random neural network* (RNN) model with the ReLU activation function. As it turns out, this hypothesis space is a convex subset of the span of a class of single-hidden-layer feed-forward networks with randomly generated coefficients. Unlike the spectral method, we determine the estimate by solving a least-squares problem. Similar to the previous section, we will focus on the issues regarding the generalization error and the Lipschitz continuity, under the same setting. The results on the approximation error and estimation error are mainly inspired by the work [30] on random neural networks and the machine-learning theory [42–44], respectively.

5.1. Hypothesis space and the approximation error

Following the notations in Section 4, we recall that $b : \mathbb{R}^d \to \mathbb{R}$ denotes a generic component of the drift coefficient, and the available i.i.d. training data $\{x_i, y_i\}_{i=1}^N \sim (X, Y)$ satisfies $y_i = b(x_i) + \eta_i$ with the Gaussian noise $\eta_i \sim E$. To introduce the hypothesis space, we define the random function $H_W^{A,\zeta} : \mathbb{R}^d \to \mathbb{R}$ by

$$H_W^{A,\zeta}(x) = \sum_{i=1}^M W_i \phi\left(\langle A_i, x\rangle + \zeta_i\right), \quad \phi(z) := \max\{0, z\}, \tag{55}$$

where $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_M)^\top \in \mathbb{R}^M$ and $A \in \mathbb{R}^{M \times d}$ (with row vectors A_i) are generated randomly. Given the realization of A and ζ, the coefficient vector $W \in \mathbb{R}^M$ is trained via an empirical risk minimization. The function ϕ in (55) is known as the ReLU activation function. We shall point out that there are other choices of activation functions. We restrict to the ReLU to directly use the results in [30] regarding the approximation error.

In particular, in [30], the approximation error is formulated with respect to a (essentially) compactly supported probability measure. To fulfill such an assumption, we introduce the following truncation to the invariant measure $\tilde{\pi}^\delta$,

$$\tilde{\pi}_D^\delta := \mathbb{1}_{B_D} \frac{\tilde{\pi}^\delta}{\tilde{\pi}^\delta(\mathbb{1}_{B_D})}, \quad B_D := \left\{x \in \mathbb{R}^d \mid \|x\| \leq D\right\}, \quad D > 1, \tag{56}$$

where $\mathbb{1}_{B_D}(\cdot)$ denotes the characteristic function with respect to the ball B_D. By the linear growth bound in Assumption 2.1, we have $|b(x)| \leq K_2 \sqrt{1 + D^2}$ for all $x \in B_D$. Fixing K_2, we introduce the following convex hypothesis space,

$$\mathscr{H}_D^{A,\zeta} := \left\{f = \mathbb{1}_{B_D} H_W^{A,\zeta} \mid \|f\|_\infty \leq K_2\sqrt{1 + D^2}\right\}.$$

Given a realization of A and ζ, we may introduce the target function, with a slight abuse of the notation,

$$b_{\mathscr{H}} := \arg\min_{h \in \mathscr{H}_D^{A,\zeta}} \|b - h\|_{\tilde{\pi}^\delta}^2 = \arg\min_{h \in \mathscr{H}_D^{A,\zeta}} \|b - h\|_{\tilde{\pi}_D^\delta}^2, \tag{57}$$

which is well-defined since $\mathscr{H}_D^{A,\zeta}$, according to the definition of the random function in (55), is a convex subset of the M-dimensional function space, $\mathrm{span}\{\phi(\langle A_i, \cdot\rangle + \zeta_i), \ i = 1, 2, \ldots, M\}$. The second identity in (57) holds because functions in the hypothesis space are supported in B_D and the measures $\tilde{\pi}^\delta$ and $\tilde{\pi}_D^\delta$ are proportional to each other in B_D.

The following result, as a direct consequence of Corollary 2 in [30], specifies how to generate the random coefficients A and ζ in the RNN (55) as well as the approximation error of $b_{\mathscr{H}}$ in (57).

Proposition 5.1. *Assume that $b \in C^k(\mathbb{R}^d)$ for some integer $k \geq \frac{d}{2} + 1 + s$ with $s > 0$. Let $T = M^{\frac{1}{2k-2s+1}}$. Suppose that the row vectors $\{A_i\}_{i=1}^M$ of the matrix A are i.i.d. samples sampled from the uniform distribution on the ball $B_T \subset \mathbb{R}^d$, and the entries of the vector ζ, $\{\zeta_i\}_{i=1}^M$, are i.i.d. samples sampled from the uniform distribution on $[-DT, DT]$. Assume the two uniform distributions together with the stationary distribution μ^δ are mutually independent. Then for any $\tau \in (0, 1)$, with probability $1 - \tau$, the target function $b_{\mathscr{H}}$ in (57) satisfies,*

$$\|b - b_{\mathscr{H}}\|_{\tilde{\pi}^\delta}^2 \leq \frac{C}{\tau} M^{-\frac{2}{\alpha}} + K_2^2 \int_{\{\|x\| > D\}} (1 + \|x\|^2)\tilde{\pi}^\delta(dx), \quad \alpha = 2 + \frac{d+1}{k - \frac{d}{2} - s}, \tag{58}$$

for some constant $C \in (0, +\infty)$, where the integral in (58) can be further bounded by

$$\int_{\{\|x\| > D\}} (1 + \|x\|^2)\tilde{\pi}^\delta(dx) \leq C_1^{-\ell}(1 + D^2)^{1-\ell}\tilde{\pi}^\delta(V). \tag{59}$$

Recall that V denotes the Lyapunov function satisfying Assumption 2.4, and the constants ℓ and C_1 are the same as in Assumption 2.4.

Proof. To begin with, since the target function $b_{\mathscr{H}}$ is supported in B_D, the L^2-error $\|b - b_{\mathscr{H}}\|_{\tilde{\pi}^\delta}^2$ is bounded as follows,

$$\|b - b_{\mathscr{H}}\|_{\tilde{\pi}^\delta}^2 = \tilde{\pi}_D^\delta(\mathbb{1}_{B_D})^2 \|b - b_{\mathscr{H}}\|_{\tilde{\pi}_D^\delta}^2 + \int_{\{\|x\| > D\}} \|b\|^2 \tilde{\pi}^\delta(dx) \leq \|b - b_{\mathscr{H}}\|_{\tilde{\pi}_D^\delta}^2 + K_2^2 \int_{\{\|x\| > D\}} (1 + \|x\|^2)\tilde{\pi}^\delta(dx),$$

where we have used the linear growth bound on b in Assumption 2.1. To apply Corollary 2 in [30] to $\|b - b_{\mathscr{H}}\|_{\tilde{\pi}_D^\delta}^2$, we introduce the mollification $b^* \in W^{k,2}(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ of b such that

$$b^* = b, \quad \forall x \in B_D.$$

Here, $W^{k,2}(\mathbb{R}^d)$ indicates a Sobolev space. We have the following relation

$$b_{\mathscr{H}} = \arg\min_{h \in \mathscr{H}_D^{A,\zeta}} \|b - h\|_{\tilde{\pi}^\delta}^2 = \arg\min_{h \in \mathscr{H}_D^{A,\zeta}} \|b - h\|_{\tilde{\pi}_D^\delta}^2 = \arg\min_{h \in \mathscr{H}_D^{A,\zeta}} \|b^* - h\|_{\tilde{\pi}_D^\delta}^2,$$

that is, $b_{\mathscr{H}}$ (57) is also the target function of b^* with respect to $L^2(\mathbb{R}^d, \tilde{\pi}_D^\delta)$. In this case, the corresponding approximation error can be bounded by Corollary 2 in [30],

$$\mathbb{E}\left[\|b - b_{\mathscr{H}}\|_{\tilde{\pi}_D^\delta}^2\right] = \mathbb{E}\left[\|b^* - b_{\mathscr{H}}\|_{\tilde{\pi}_D^\delta}^2\right] \leq C M^{-\frac{2}{\alpha}}, \quad \alpha = 2 + \frac{d+1}{k - \frac{d}{2} - s},$$

for some constant $C \in (0, +\infty)$. Here the expectation is taken with respect to the random coefficients A and ζ. Together with the Markov's inequality, we reach the upper bound in (58).

As for the integral in (58), simply notice that by Assumption 2.4, one has,

$$1 + \|x\|^2 = \frac{(1 + \|x\|^2)^\ell}{(1 + \|x\|^2)^{\ell-1}} \le C_1^{-\ell}(1 + D^2)^{1-\ell}V(x), \quad \forall x \in \{\|x\| > D\},$$

which leads to the upper bound in (59). □

We want to point out that in [30], an explicit expression of the constant C in (58) has been provided, which is proportional to the square of the norm of b^* in $W^{k,2}(\mathbb{R}^d)$.

Remark 2. We should also point out that the polynomial decay bound in (59) can be improved under additional assumptions. Particularly, if X is a centered random variable of sub-exponential distribution $SE(\nu^2, \alpha)$ with $\nu, \alpha > 0$, then by concentration inequality for sub-exponential distribution, one obtains

$$\mathbb{P}(\|X\| \ge D) \le e^{-\frac{D}{2\alpha}}, \quad \forall D > \nu^2\alpha^{-1}.$$

This means,

$$\int_{\{\|x\|>D\}} (1 + \|x\|^2)\tilde{\pi}^\delta(dx) \le \mathbb{P}(\|X\| \ge D)^{1/2}\left(\int_{\mathbb{R}^d} (1 + \|x\|^2)^2\tilde{\pi}^\delta(dx)\right)^{1/2} \le e^{-\frac{D}{4\alpha}}\tilde{\pi}^\delta(V^2)^{1/2}, \tag{60}$$

decays exponentially in D when $\tilde{\pi}^\delta(V^2) < \infty$.

5.2. The generalization error

Due to the truncation to the distribution $\tilde{\pi}^\delta$ in (56), we shall define the risk function $\mathscr{E}[h]$ in (18) with respect to the truncated random variables (X_D, Y_D), where X_D follows the distribution $\tilde{\pi}_D^\delta$ and Y_D is still determined by the model (16). Let μ_D^δ denote their joint distribution, and the risk function in (18) becomes

$$\mathscr{E}_D[h] := \mathbb{E}_{\mu_D^\delta}\left[\|h(X_D) - Y_D\|^2\right].$$

For simplicity, we assume the constant D is large enough so that the samples $\{x_i\}_{i=1}^N \subset B_D$. Then, the corresponding empirical risk function, denoted as $\mathscr{E}_{D,N}$, is the same as the risk $\mathscr{E}_N$ in (19). In particular, following (20), the empirical estimate $b_{\epsilon,D}$ is given by,

$$b_{\epsilon,D} := \arg\min_{h \in \mathscr{H}_D^{A,\zeta}} \mathscr{E}_{D,N}[h]. \tag{61}$$

The least-squares problem in (61) is conditionally linear in the sense that, given a realization of A and ζ, functions in $\mathscr{H}_D^{A,\zeta}$ are linear combinations of basis functions $\{\phi(\langle A_i, \cdot\rangle + \zeta_i)\}_{i=1}^M$. Thus, if we let

$$b_{\epsilon,D}(x) = \mathbb{1}_{B_D}(x)\sum_{i=1}^M \hat{W}_i\phi(\langle A_i, x\rangle + \zeta_i), \tag{62}$$

then $\hat{W} = (\hat{W}_1, \hat{W}_2, \ldots, \hat{W}_M)^\top$ solves the linear system

$$\Phi^\top\Phi\hat{W} = \Phi^\top y, \quad y = (y_1, y_2, \ldots, y_N)^\top, \tag{63}$$

where $\Phi \in \mathbb{R}^{N \times M}$ such that

$$\Phi = (\Phi_{ij}), \quad \Phi_{ij} = \phi(\langle A_j, x_i\rangle + \zeta_j), \quad 1 \le i \le N, \quad 1 \le j \le M.$$

In practice, when $\Phi^\top\Phi$ in (63) is singular, we write $\hat{W} = (\Phi^\top\Phi)^\dagger\Phi^\top y$, with $(\Phi^\top\Phi)^\dagger$ being the Moore–Penrose inverse.

Various results on the estimation error of empirical risk minimization are available under a bounded sampling assumption, e.g., [44]. However, due to the Gaussian noise η_i in y_i, our problem belongs to the unbounded sampling case. As a remedy, we will consider the result in [43] to study the estimation error $\mathscr{E}_D[b_{\epsilon,D}] - \mathscr{E}_D[b_{\mathscr{H}}]$.

Proposition 5.2. Let $\{x_i, y_i\}_{i=1}^N$ be i.i.d. samples with $x_i \sim X_D$ of distribution $\tilde{\pi}_D^\delta$ and $y_i = b(x_i) + \eta_i$ with Gaussian noise $\eta_i \sim E = \mathcal{N}(0, R)$ independent of $\tilde{\pi}_D^\delta$. Then, for any $\tau \in (0, 1)$, with probability $1 - \tau$, the estimation error satisfies

$$\mathscr{E}_D[b_{\epsilon,D}] - \mathscr{E}_D[b_{\mathscr{H}}] \le \frac{B_1}{\sqrt{N}}\left(\ln\frac{8}{\tau^2} + M\ln\ln\frac{2}{\tau}\right) + \frac{B_1 M}{\sqrt{N}}\left(1 + \ln\left(\frac{4K_2 B_2\sqrt{1 + D^2}}{B_1} \cdot \frac{\sqrt{N}}{M}\right)\right),$$

for some constants $B_1, B_2 \in (0, +\infty)$, which will be specified in the proof.

Proof. The result in [43] relies on the notion of *covering number of* the hypothesis space $\mathscr{H}_D^{A,\zeta}$ and a moment condition on the output y_i.

Recall that for $\theta > 0$, the covering number of $\mathscr{H}_D^{A,\zeta}$ with radius θ, denoted by $N(\mathscr{H}_D^{A,\zeta}, \theta)$, is defined to be the minimal integer n such that there exist n balls with radius θ covering $\mathscr{H}_D^{A,\zeta}$. Here, the topology are induced by the uniform norm $\|\cdot\|_\infty$. Notice that $\mathscr{H}_D^{A,\zeta}$ is a bounded subset of an M-dimensional normed space. By Proposition 5 in [44], we have

$$\ln N(\mathscr{H}_D^{A,\zeta}, \theta) \le M\ln\left(\frac{4K_2\sqrt{1 + D^2}}{\theta}\right), \quad \forall\theta > 0.$$

The output y_i yields a Gaussian distribution $\rho(y|x_i) = \mathcal{N}(b(x_i), R)$. The moment of the output $\mathbb{E}[|b(X_D) + E|^n]$ is given by the following integral

$$\mathbb{E}\left[|b(X_D) + E|^n\right] = \int_{B_D} \int_{\mathbb{R}} |y|^n \rho(y|x) dy \tilde{\pi}_D^\delta(dx), \quad n \in \mathbb{N},$$

where for all $x \in B_D$,

$$\int_{\mathbb{R}} |y|^n \rho(y|x) dy \leq 2^{n-1} \left(\int_{\mathbb{R}} |y - b(x)|^n \rho(y|x) dy + \int_{\mathbb{R}} |b(x)|^n \rho(y|x) dy \right) \leq 2^{n-1} \left((2R)^{\frac{n}{2}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}} + K_2^n (1 + D^2)^{\frac{n}{2}} \right).$$

Here, we have used the identity regarding the central absolute moments of Gaussian distribution and the linear growth bound of b. Since the Gamma function is monotone increasing, that is,

$$\Gamma\left(\frac{n+1}{2}\right) \leq \Gamma\left(m + \frac{1}{2}\right) = \frac{(2m-1)!!}{2^m} \sqrt{\pi} \leq \frac{n!!}{2^{\frac{n}{2}}} \sqrt{\pi}, \quad m = \lceil \frac{n}{2} \rceil$$

($\lceil \frac{n}{2} \rceil$ denotes the smallest integer that is greater or equal than $\frac{n}{2}$), which suggests

$$(2R)^{\frac{n}{2}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}} \leq R^{\frac{n}{2}} n!! \leq R^{\frac{n}{2}} \sqrt{n!}.$$

Choosing,

$$M_D := 2 \max\left\{ K_2 \sqrt{1 + D^2}, \sqrt{R} \right\}, \tag{64}$$

we obtain,

$$\mathbb{E}\left[|b(X_D) + E|^n\right] \leq \sqrt{n!} M_D^n, \quad \forall n \in \mathbb{N}, \tag{65}$$

which fulfills the moment hypothesis in [43].

By Propositions 2.2 and 3.3 in [43], we see that, with probability at least $1 - \tau$, there holds for all $\theta > 0$

$$\mathscr{E}_D[b_{\epsilon,D}] - \mathscr{E}_D[b_{\mathscr{H}}] \leq \left(\mathscr{E}_D[b_{\epsilon,D}] - \mathscr{E}_{D,N}[b_{\epsilon,D}] \right) + \left(\mathscr{E}_{D,N}[b_{\mathscr{H}}] - \mathscr{E}_D[b_{\mathscr{H}}] \right)$$

$$\leq \frac{B_1}{\sqrt{N}} \left(\ln \frac{8}{\tau^2} + \ln N(\mathscr{H}_D^{A,\zeta}, \theta) \right) + B_2 \theta \ln \frac{2}{\tau}$$

$$\leq \frac{B_1}{\sqrt{N}} \ln \frac{8}{\tau^2} + \left(\frac{B_1 M}{\sqrt{N}} \ln \left(\frac{4K_2\sqrt{1 + D^2}}{\theta} \right) + B_2 \theta \ln \frac{2}{\tau} \right),$$

where the constants B_1 and B_2 are given by

$$B_1 = 40 K_2^2 (1 + D^2) + 160 M_D^2 \leq 180 M_D^2, \quad B_2 = 4\left[K_2(1 + D^2)^{\frac{1}{2}} + (2 + 3\sqrt{2})M_D \right] \leq (10 + 12\sqrt{2})M_D. \tag{66}$$

In particular, when

$$\theta = \theta^* := \frac{B_1 M}{B_2 \ln \frac{2}{\tau} \sqrt{N}} > 0$$

the upper bound reaches its minimum, and we have

$$\mathscr{E}_D[b_{\epsilon,D}] - \mathscr{E}_D[b_{\mathscr{H}}] \leq \frac{B_1}{\sqrt{N}} \ln \frac{8}{\tau^2} + \frac{B_1 M}{\sqrt{N}} \left[1 + \ln \left(\frac{4K_2 B_2 \sqrt{1 + D^2}}{B_1} \right) + \ln \ln \frac{2}{\tau} + \ln \frac{\sqrt{N}}{M} \right]. \quad \square$$

It is worthwhile mentioning that the moment condition in (65) is a necessary condition for using the Bennet inequality to deduce the bounds above. Importantly, this moment bound gives a characterization of the estimation error in terms of the noise variance R through constant M_D in (64) that appears in both B_1 and B_2 as shown in (66). Recall that in our application $R \leq \delta^{-1} Tr[\sigma\sigma^\top]$ as in (46). Using the covering number of the hypothesis space to analyze the estimation error is a classical approach in statistical learning theory, e.g., [66]. Combining Propositions 5.1–5.2, with probability $1 - \tau$, we conclude the following upper bound for the generalization error,

$$\mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_{\epsilon,D}\|^2 \right] = \mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_{\mathscr{H}}\|^2 \right] + \left(\mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_{\epsilon,D}\|^2 \right] - \mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_{\mathscr{H}}\|^2 \right] \right)$$

$$\leq \mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_{\mathscr{H}}\|^2 \right] + \left(\mathbb{E}_{\tilde{\pi}_D^\delta}\left[\|b - b_{\epsilon,D}\|^2 \right] - \mathbb{E}_{\tilde{\pi}_D^\delta}\left[\|b - b_{\mathscr{H}}\|^2 \right] \right)$$

$$= \mathbb{E}_{\tilde{\pi}^\delta}\left[\|b - b_{\mathscr{H}}\|^2 \right] + \mathscr{E}_D\left[b_{\epsilon,D} \right] - \mathscr{E}_D\left[b_{\mathscr{H}} \right]$$

$$= O\left(\tau^{-1} M^{-\frac{2}{\alpha}} \right) + O\left((1 + D^2)^{1-\ell} \right) + O\left(M_D^2 N^{-\frac{1}{2}} \left(\ln \tau^{-2} + M \ln \ln \tau^{-1} \right) \right) + O\left(M_D^2 \frac{M}{\sqrt{N}} \ln \frac{\sqrt{N}}{M} \right),$$

where the constant M_D depends on the noise variance as defined in (64). If we assume the invariant measure $\tilde{\pi}^\delta$ yields an exponential decay as in Remark 2, then the second term above can be replaced by the error bound in (60). For fixed $\tau > 0$, the leading error term is the last component. Choosing $N = O(M^2)$, the last error term is $O(M_D^2)$ which is effectively $O(R)$, where $R \leq \delta^{-1} Tr[\sigma\sigma^\top]$. This means the contribution from noise is comparable to that in the error from the kernel method in (52).

5.3. Lipschitz continuity

To close this section, we check the Lipschitz continuity of the estimator $b_{\epsilon,D}$ in (61). From (62), we have

$$\left| b_{\epsilon,D}(x) - b_{\epsilon,D}(y) \right| \leq \sum_{i=1}^{M} |\hat{W}_i| \, |\phi\left(\langle A_i, x\rangle + \zeta_i\right) - \phi\left(\langle A_i, y\rangle + \zeta_i\right)| \leq \sum_{i=1}^{M} |\hat{W}_i| \|A_i\| \|x - y\|$$

$$\leq \left(\sum_{i=1}^{M} \hat{W}_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{M} \|A_i\|^2\right)^{\frac{1}{2}} \|x - y\| = \|\hat{W}\| \|\boldsymbol{A}\|_F \|x - y\|,$$

which suggests that $b_{\epsilon,D}$ is globally Lipschitz. Recall that the matrix $\boldsymbol{A}$ is randomly generated following Proposition 5.1. We have the following concentration bounds for $\|\boldsymbol{A}\|_F$.

Proposition 5.3. Let $\boldsymbol{A} \in \mathbb{R}^{M \times d}$ with row vectors A_i randomly generated as in Proposition 5.1. Then, for $\tau > 0$ we have

$$\mathbb{P}\left(\|\boldsymbol{A}\|_F^2 \in [M\mu_A - \tau, M\mu_A + \tau]\right) \geq 1 - 2\exp\left(-\frac{\tau^2}{2M\sigma_A^2 + \frac{2}{3}\mu_A\tau}\right), \quad \mu_A = \frac{d}{d+2}T^2, \quad \sigma_A^2 = \frac{4d}{(d+4)(d+2)^2}T^4,$$

where the constant T is the same as in Proposition 5.1.

Proof. Recall that the row vectors $\{A_i\}_{i=1}^{M}$ are i.i.d. samples drawn from the uniform distribution on the ball $B_T \subset \mathbb{R}^d$. Let ξ denote the random variable corresponding to $\|A_i\|$ with density ρ_T. Using the spherical coordinates, one sees that $\rho_T(r) \propto r^{d-1}$. Moreover, ρ_T is supported in $[0, T]$, that is,

$$\rho_T(r) = \frac{d}{T^d} r^{d-1}, \quad r \in [0, T].$$

Since $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{M} \|A_i\|^2$, to apply concentration inequalities, we need to identify the statistics of the random variable ξ^2. By direct computation, we find

$$\mu_A := \mathbb{E}[\xi^2] = \frac{d}{d+2}T^2, \quad \sigma_A^2 := \mathbb{E}[\xi^4] - \mathbb{E}[\xi^2]^2 = \frac{4d}{(d+4)(d+2)^2}T^4.$$

Further notice that $\xi^2 \in [0, T^2]$ and $|\xi^2 - \mu_A| \leq \mu_A$, and by the Bernstein inequality (e.g., Proposition 2 in [44]), we reach the bound in the proposition's statement. $\square$

We want to point out that since $T \leq M^{\frac{1}{d+3}}$ according to Proposition 5.1, Proposition 5.3 suggests that

$$\|\boldsymbol{A}\|_F = O\left(M^{\frac{d+5}{2d+6}}\right),$$

in high probability.

As for $\|\hat{W}\|$, since $\hat{W}$ solves the linear system in (63) the norm $\|\hat{W}\|$ depends on the smallest positive eigenvalue of the matrix $\Phi^\top \Phi$. Thus, in practice, to control $\|\hat{W}\|$, a regularization is necessary. This provably improves the Lipschitz continuity of the estimates, but it comes in the expense of estimation error. For general discussions on the Lipschitz function approximation using neural network, we refer the readers to [67] and the references therein.

5.4. Remarks on the ReLU random neural networks

The ReLU RNN approach reviewed here has several advantages. First, compared with the spectral regression method in Section 4, the RNN approach is numerically cheaper to implement since it does not require solving large eigenvalue problems. Given the realization of $\boldsymbol{A}$ and ζ, the target function $b_{\mathscr{H}}$ can be properly defined as the minimizer of the least-squares problem in (57) with no extra assumption on the unknown function b. The existence of a target function simplifies the discussion of the generalization error.

An obvious drawback of the RNN approach is that the estimates $b_{\epsilon,D}$ (62) has compact support, while the unknown drift coefficients $b(x)$ are often unbounded as $\|x\| \to +\infty$. Under such constructions, the consistency condition (25) in Assumption 3.1 can never be satisfied beyond the compact support. While the error beyond the compact support decays, either polynomial (see (59)) or exponential (see (60)) under additional assumptions, as a function of the radius of the ball, in practice, we may not be able to estimate on a domain with large D. This issue is due to the difficulty in obtaining training samples on the tail of the distribution X. Additionally, larger D induces a larger estimation error through the constant M_D in (64).

6. Summary

In this paper, we studied the error bounds of the invariant statistics in learning ergodic Itô diffusion. Using the perturbation theory of ergodic Markov chains [37,38] and the linear response theory [40], we established a linear dependence of the errors of one-point and two-point invariant statistics on the spectral error of the diffusion matrix estimator. Under a proper consistency condition on the estimator of the drift coefficient, one can identify the error bound in terms of the size of the training sample, "size" of hypothesis space, noise amplitude, and the discretization error induced by the SDE's solver, using standard L^2 generalization error analysis corresponding to the specific machine learning algorithm. An important takeaway from this study is that the L^2 characterization of the learning generalization error is not sufficient for achieving the linear dependence error bound presented in this paper, due to the discrepancies

of the topologies that characterize the errors in perturbation theory of ergodic Markov chain and learning theory. Besides the consistent estimator in the hypothesis space that retains certain characteristics of the drift coefficients, a sufficient condition to achieve the error bound here is through learning algorithms that produce a uniformly Lipschitz estimator.

From our examination of two simple learning algorithms, the kernel-based spectral regression method and the shallow random neural networks with the ReLU activation function, we conclude that to satisfy these sufficient conditions, one needs to specify the hypothesis space carefully to avoid bias. In practice, the main challenge will be in the kernel specification when a kernel-based algorithm is used. For random neural network, the consistency depends on the universality of the random bases [26]. Another practical issue is to overcome biased estimation with the random neural networks that arises when sampling data on the tail of the distribution are not available. These issues suggest that it is important to have a physical understanding of the problem at hand to have appropriate hypothesis space for convergence guarantees rather than just using machine learning as a black-box. Furthermore, consistent estimates can only be achieved on the domain where the training data is available.

We view this study as a first step to understand machine learning of dynamical systems with invariant statistical properties. Many issues remain open. For example, in our study, we consider a consistent supervised learning problem in the following sense. Specifically, we model the response variable y in (16) to be compatible with the integration Euler–Maruyama scheme. In practice, when the underlying scheme is not known, the inconsistent learning model will produce a biased estimator. Another related issue is the global Lipschitz condition on the drift term. If we relax it to be locally Lipschitz, the EM discretization does not yield an ergodic Markov chain in general. As pointed out in [39], one can generate an ergodic Markov chain by considering the stochastic backward Euler discretization. With this discretization, unfortunately, the corresponding supervised learning model will be much more complicated to analyze since the loss function is implicit. If one uses the learning model corresponding to EM discretization or any other explicit integration schemes, then an additional bias will be introduced. A much harder yet important problem is to carry this analysis on deterministic dynamical systems. In this context, the existence of the invariant measure of the estimated dynamics is an essential question in dynamical system theory [68]. Furthermore, the validity of linear response theory is also a critical problem [69] that is difficult to justify in general.

While this paper focuses on the theory, several numerical results using kernel methods are readily available. For example, see Section 3 of [14] for numerical results on a linear SDE and Section 2.2 of [19] for numerical results on a nonlinear Langevin equation. Motivated by the analysis in this paper, we are pursuing numerical studies with neural network models in separate works.

CRediT authorship contribution statement

He Zhang: Conceptualization, Investigation, Formal analysis, Writing – original draft. **John Harlim:** Conceptualization, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Xiantao Li:** Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Proof of Lemma 3.1

In this appendix, we discuss the proof of Lemma 3.1 in Section 3. For reader's convenience, we first review a nonlinear generalization of Gronwall's inequality.

Proposition A.1. *Let $v(t)$ be a nonnegative function that satisfies the integral inequality*

$$v(t) \le c + \int_{t_0}^{t} (a_1(s)v(s) + a_2(s)v^\alpha(s)) \, ds, \quad c \ge 0, \quad \alpha \ge 0,$$

where $a_1(t)$ and $a_2(t)$ are continuous nonnegative functions on $[t_0, +\infty)$. For $0 \le \alpha < 1$, we have

$$v(t) \le \left\{ c^{1-\alpha} \exp\left[(1-\alpha) \int_{t_0}^{t} a_1(s)ds \right] + (1-\alpha) \int_{t_0}^{t} a_2(s) \exp\left[(1-\alpha) \int_{s}^{t} a_1(r)dr \right] ds \right\}^{\frac{1}{1-\alpha}}.$$

When $\alpha = 1$, such a result reduces to the standard Gronwall's inequality. The proof is an application of the Bernoulli equation (e.g., Theorem 21 in [70]).

Proof of Lemma 3.1. We start with the case: $(u_n, u_n^\epsilon) = (X_n, X_n^\epsilon)$. By the definition of $\mathscr{G}_\ell$ in Proposition 2.2, we have

$$|f(x) - f(y)| \le C_\ell \left(1 + \|x\|^{2\ell-1} + \|y\|^{2\ell-1} \right) \|x - y\|, \quad \forall x, y \in \mathbb{R}^d, \ \forall f \in \mathscr{G}_\ell, \tag{67}$$

where $C_\ell \in (0, +\infty)$ is a fixed constant independent of f. Taking expectation on (67), and employing the Cauchy–Schwarz inequality, we obtain

$$\left| \mathbb{E}^x[f(X_1)] - \mathbb{E}^x[f(X_1^\epsilon)] \right| \le C_\ell \left(\mathbb{E}^x \left[\left(1 + \|X_1\|^{2\ell-1} + \|X_1^\epsilon\|^{2\ell-1} \right)^2 \right] \right)^{\frac{1}{2}} \left(\mathbb{E}^x \left[\|X_1 - X_1^\epsilon\|^2 \right] \right)^{\frac{1}{2}},$$

Further notice that

$$\mathbb{E}^x\left[\left(1 + \|X_1\|^{2\ell-1} + \|X_1^\epsilon\|^{2\ell-1}\right)^2\right] = \mathbb{E}^x\left[1 + \|X_1\|^{4\ell-2} + \|X_1^\epsilon\|^{4\ell-2}\right] + 2\mathbb{E}^x\left[\|X_1\|^{2\ell-1} + \|X_1^\epsilon\|^{2\ell-1} + \|X_1\|^{2\ell-1}\|X_1^\epsilon\|^{2\ell-1}\right],$$

where the last term can be bounded as follows,

$$\mathbb{E}^x\left[\|X_1\|^{2\ell-1}\|X_1^\epsilon\|^{2\ell-1}\right] \leq \left(\mathbb{E}^x\left[\|X_1\|^{4\ell-2}\right]\right)^{\frac{1}{2}}\left(\mathbb{E}^x\left[\|X_1^\epsilon\|^{4\ell-2}\right]\right)^{\frac{1}{2}} \leq \frac{1}{2}\left(\mathbb{E}^x\left[\|X_1\|^{4\ell-2}\right] + \mathbb{E}^x\left[\|X_1^\epsilon\|^{4\ell-2}\right]\right).$$

To bound the remaining order-$(2\ell-1)$ moments of X_1 and X_1^ϵ, we apply the Jensen's inequality,

$$\mathbb{E}^x\left[\|A\|^{2\ell-1}\right] \leq \mathbb{E}^x\left[\|A\|^{4\ell-2}\right]^{\frac{1}{2}} \leq \max\left\{\mathbb{E}^x\left[\|A\|^{4\ell-2}\right], 1\right\},$$

(for $A = X_1$ and X_1^ϵ, respectively), which leads to

$$\left|\mathbb{E}^x[f(X_1)] - \mathbb{E}^x[f(X_1^\epsilon)]\right| \leq R_1\left(\mathbb{E}^x\left[1 + \|X_1\|^{4\ell-2} + \|X_1^\epsilon\|^{4\ell-2}\right]\right)^{\frac{1}{2}}\left(\mathbb{E}^x\left[\|X_1 - X_1^\epsilon\|^2\right]\right)^{\frac{1}{2}} \tag{68}$$

for some constant $R_1 \in (0, +\infty)$ independent of ϵ. On the right-hand side of Eq. (68), the moments are bounded by applying Lemma 2.1 to $X(t)$ and $X^\epsilon(t)$, respectively. As for the error $\mathbb{E}^x\left[\|X_1 - X_1^\epsilon\|^2\right]$ in (68), we will derive an integral inequality of the form in Proposition A.1 using Itô formula [2].

Notice that since $X(t)$ in (9) and $X^\epsilon(t)$ in (15) are driven by the same Brownian motion W, we can consider the following coupled SDEs

$$\frac{d}{dt}\begin{pmatrix} X \\ X^\epsilon \end{pmatrix} = \begin{pmatrix} b(X) \\ b_\epsilon(X^\epsilon) \end{pmatrix} + \begin{pmatrix} \sigma \\ \sigma_\epsilon \end{pmatrix}\dot{W}, \quad \begin{pmatrix} X(0) \\ X^\epsilon(0) \end{pmatrix} = \begin{pmatrix} x \\ x \end{pmatrix}. \tag{69}$$

Recall that in Section 3.1, we estimated $\sigma\sigma^\top$ by $\sigma_\epsilon\sigma_\epsilon^\top$ without specifying the diffusion coefficient σ_ϵ. Here, we shall take σ_ϵ so that $\|\sigma - \sigma_\epsilon\|_F = O(\epsilon)$ for the sake of the proof. To construct such σ_ϵ, we introduce the *thin* SVD factorization of σ (recall that $\sigma \in \mathbb{R}^{d\times m}$ is full rank with $m \leq d$),

$$\sigma = U\Lambda V^\top, \quad U \in \mathbb{R}^{d\times m}, \quad \Lambda \in \mathbb{R}^{m\times m}, \quad V \in \mathbb{R}^{m\times m}, \tag{70}$$

where $\Lambda = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m)$ with $\sigma_i > 0, \forall i$. With the SVD factorization, we get $\sigma\sigma^\top = U\Lambda^2 U^\top$. Given the estimate $\sigma_\epsilon\sigma_\epsilon^\top$ of $\sigma\sigma^\top$ satisfying (24), we define

$$\Sigma_\epsilon := U^\top \sigma_\epsilon\sigma_\epsilon^\top U \in \mathbb{R}^{m\times m},$$

and we have

$$\|\Lambda^2 - \Sigma_\epsilon\|_2 = \|U^\top(\sigma\sigma^\top - \sigma_\epsilon\sigma_\epsilon^\top)U\|_2 = \|\sigma\sigma^\top - \sigma_\epsilon\sigma_\epsilon^\top\|_2 = \epsilon. \tag{71}$$

Since $\Lambda^2 = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ is positive definite, for ϵ small enough, Σ_ϵ is also positive definite according to (71). Thus, there exists a unique lower triangular Cholesky factorization of Σ_ϵ, e.g., Theorem 4.2.7 in [71], with the lower triangular matrix denoted by $L_\epsilon \in \mathbb{R}^{m\times m}$. Namely,

$$L_\epsilon L_\epsilon^\top = \Sigma_\epsilon.$$

Moreover, by the forward stability of the Cholesky factorization subject to small perturbations, e.g., Theorem 2.1 in [72], we have

$$\|\Lambda - L_\epsilon\|_F \leq \|\Lambda^2 - \Sigma_\epsilon\|_F \leq \sqrt{m}\|\Lambda^2 - \Sigma_\epsilon\|_2 = \sqrt{m}\epsilon. \tag{72}$$

In other words, L_ϵ approximates the matrix Λ in the SVD factorization of σ (70). Thus, by replacing Λ in (70) by L_ϵ, we define

$$\sigma_\epsilon := UL_\epsilon V^\top, \tag{73}$$

which satisfies

$$\left(UL_\epsilon V^\top\right)\left(UL_\epsilon V^\top\right)^\top = U\Sigma_\epsilon U^\top = \sigma_\epsilon\sigma_\epsilon^\top.$$

In particular, by (72), we have

$$\|\sigma - \sigma_\epsilon\|_F \leq \sqrt{m}\|\sigma - \sigma_\epsilon\|_2 = \sqrt{m}\|U(\Lambda - L_\epsilon)V^\top\|_2 \leq \sqrt{m}\|\Lambda - L_\epsilon\|_2 \leq m\epsilon. \tag{74}$$

We assign σ_ϵ in (73) to the coupled system (69), and apply the Itô formula to the process

$$U(t) := \|X(t) - X^\epsilon(t)\|^2, \quad U(0) = 0, \quad 0 \leq t \leq \delta.$$

Direct calculations yield,

$$\dot{U} = 2\langle X - X^\epsilon, b(X) - b_\epsilon(X^\epsilon)\rangle + 2\langle X - X^\epsilon, (\sigma - \sigma_\epsilon)\dot{W}\rangle + \langle(\sigma - \sigma_\epsilon)\dot{W}, (\sigma - \sigma_\epsilon)\dot{W}\rangle$$

$$= 2\langle X - X^\epsilon, b(X) - b_\epsilon(X^\epsilon)\rangle + \|\sigma - \sigma_\epsilon\|_F^2 + 2\langle X - X^\epsilon, (\sigma - \sigma_\epsilon)\dot{W}\rangle,$$

where $\langle\cdot, \cdot\rangle$ denotes the inner product in $\mathbb{R}^d$. This can be rewritten as an Itô integral representation of $U(t)$,

$$U(t) = \int_0^t 2\langle X(s) - X^\epsilon(s), b(X(s)) - b_\epsilon(X^\epsilon(s))\rangle + \|\sigma - \sigma_\epsilon\|_F^2 ds + \int_0^t (\sigma - \sigma_\epsilon)^\top(X - X^\epsilon) \cdot dW.$$

Here,

$$\int_0^t (\sigma - \sigma_\epsilon)^\top(X - X^\epsilon) \cdot dW := \sum_{i=1}^m \int_0^t (\sigma - \sigma_\epsilon)_i^\top(X - X^\epsilon)dW_i,$$

where $(\sigma - \sigma_\epsilon)_i \in \mathbb{R}^d$ denotes the ith column vector of the matrix $(\sigma - \sigma_\epsilon)$. Using the Itô-isometry [2], we have

$$\mathbb{E}^x \left[\int_0^t (\sigma - \sigma_\epsilon)^\top (X - X^\epsilon) \cdot dW \right] \leq \mathbb{E}^x \left[\int_0^t \|(\sigma - \sigma_\epsilon)^\top (X - X^\epsilon)\|^2 ds \right]^{\frac{1}{2}} \leq \|\sigma - \sigma_\epsilon\|_F \int_0^t \mathbb{E}^x[U(s)]^{\frac{1}{2}} ds.$$

Further observe that $b(X) - b_\epsilon(X^\epsilon) = (b(X) - b(X^\epsilon)) + (b(X^\epsilon) - b_\epsilon(X^\epsilon))$. By the Lipschitz continuity of b in Assumption 2.1 and the linear growth bound of $b - b_\epsilon$ in Assumption 3.1, we have

$$|\langle X - X^\epsilon, b(X) - b_\epsilon(X^\epsilon) \rangle| \leq |\langle X - X^\epsilon, b(X) - b(X^\epsilon) \rangle| + |\langle X - X^\epsilon, b(X^\epsilon) - b_\epsilon(X^\epsilon) \rangle|$$

$$\leq K_1 \|X - X^\epsilon\|^2 + \epsilon K_3 (1 + \|X^\epsilon\|^2)^{\frac{1}{2}} \|X - X^\epsilon\|.$$

Combining these inequalities, we arrive at,

$$\mathbb{E}^x[U(t)] \leq t\|\sigma - \sigma_\epsilon\|_F^2 + 2 \int_0^t K_1 \mathbb{E}^x[U(s)] + \left(\epsilon K_3 \mathbb{E}^x \left[(1 + \|X^\epsilon(s)\|^2)^{\frac{1}{2}} \right] + \|\sigma - \sigma_\epsilon\|_F \right) \mathbb{E}^x[U(s)]^{\frac{1}{2}} ds, \quad \forall t \in [0, \delta].$$

Thus, the function $v(t; x) := \mathbb{E}^x[U(t)]$ satisfies the following integral inequality

$$v(t; x) \leq c + \int_0^t \left(a_1 v(s) + a_2 v^{\frac{1}{2}}(s) \right) ds, \quad \forall t \in (0, \delta),$$

with

$$c = \delta m^2 \epsilon^2, \quad a_1 = 2K_1, \quad a_2 = R_2(1 + \|x\|^2)^{\frac{1}{2}} \epsilon,$$

where the constant $R_2 \in (0, +\infty)$ is independent of ϵ. Here, we have used the error bound on $\|\sigma - \sigma_\epsilon\|_F$ in Eq. (74) to get c and the moment bound of $\mathbb{E}^x \left[\|X^\epsilon\|^2 \right]$ (by applying Lemma 2.1 to X^ϵ) to get a_2. By the Gronwall-type inequality in Proposition A.1, we conclude the following bound for v,

$$\mathbb{E}^x \left[\|X(t) - X^\epsilon(t)\|^2 \right] = v(t; x) \leq R_3 e^{2K_1 t}(1 + \|x\|^2)\epsilon^2, \quad \forall t \in [0, \delta],$$

where the constant $R_3 \in (0, +\infty)$ is independent of ϵ. Thus, Eq. (68) becomes

$$\left| \mathbb{E}^x[f(X_1)] - \mathbb{E}^x[f(X_1^\epsilon)] \right| \leq R_1 R_3^{\frac{1}{2}} e^{K_1 \delta}(1 + \|x\|^2)^{\frac{1}{2}} \left(\mathbb{E}^x \left[1 + \|X_1\|^{4\ell-2} + \|X_1^\epsilon\|^{4\ell-2} \right] \right)^{\frac{1}{2}} \epsilon \leq KV(x)\epsilon,$$

for some constant $K \in (0, +\infty)$ independent of ϵ. In the last inequality, we have used the moment bounds on X_1 and X_1^ϵ based on Lemma 2.1 and Assumption 2.4 on the Lyapunov function V. Finally, by the definition of γ_X, we have,

$$\gamma_X = \sup_{x \in \mathbb{R}^d} \sup_{f \in \mathscr{G}_\ell} \frac{\left| \mathbb{E}^x[f(X_1)] - \mathbb{E}^x[f(X_1^\epsilon)] \right|}{V(x)} \leq K\epsilon,$$

which is the desirable result.

When $(u_n, u_n^\epsilon) = (X_n^\delta, X_n^{\epsilon, \delta})$, as an analogy of (68), we have

$$\left| \mathbb{E}^x[f(X_1^\delta)] - \mathbb{E}^x[f(X_1^{\epsilon, \delta})] \right| \leq R_1 \left(\mathbb{E}^x \left[1 + \|X_1^\delta\|^{4\ell-2} + \|X_1^{\epsilon, \delta}\|^{4\ell-2} \right] \right)^{\frac{1}{2}} \left(\mathbb{E}^x \left[\|X_1^\delta - X_1^{\epsilon, \delta}\|^2 \right] \right)^{\frac{1}{2}}.$$

The bound for $\mathbb{E}^x \left[\|X_1^\delta - X_1^{\epsilon, \delta}\|^2 \right]$ can be derived via direct computations. We apply EM scheme (7) to the coupled system (69), and we have

$$\mathbb{E}^x \left[\|X_1^\delta - X_1^{\epsilon, \delta}\|^2 \right] = \mathbb{E} \left[\|\delta(b(x) - b_\epsilon(x)) + \sqrt{\delta}(\sigma - \sigma_\epsilon)\xi_1\|^2 \right]$$

$$= \delta^2 \|b(x) - b_\epsilon(x)\|^2 + \delta \mathbb{E} \left[\|(\sigma - \sigma_\epsilon)\xi_1\|^2 \right] \leq R_4 (1 + \|x\|^2)\epsilon^2,$$

for some constant $R_4 \in (0, +\infty)$ independent of ϵ. Repeating the argument for γ_X, we reach the same desirable result for γ_{X^δ}.

Appendix B. Proof of Proposition 3.2

In this Appendix, we discuss the proof of Proposition 3.2, which provides both the well-posedness of the two-point statistics of the perturbed dynamics and the corresponding error bound.

By the finite second moments assumption, we know the two-point statistics in (30) are well-defined. In particular, we have

$$(k_{A,B}^\epsilon)_n - (k_{A,B})_n = \iint A(x)B(x_0)P_{n\delta}^\epsilon(x_0, dx)\pi^\epsilon(dx_0) - \iint A(x)B(x_0)P_{n\delta}(x_0, dx)\pi(dx_0)$$

$$= \iint A(x)B(x_0) \left(P_{n\delta}^\epsilon(x_0, dx) - P_{n\delta}(x_0, dx) \right) \pi^\epsilon(dx_0) + \iint A(x)B(x_0)P_{n\delta}(x_0, dx)(\pi^\epsilon - \pi)(dx_0) =: I_1 + I_2.$$

For the term I_1, we have

$$|I_1| \leq \int |B(x_0)| \left| \int A(x) \left(P_{n\delta}^\epsilon(x_0, dx) - P_{n\delta}(x_0, dx) \right) \right| \pi^\epsilon(dx_0) = \int |B(x_0)| \left| \mathbb{E}^{x_0}[A(X_n^\epsilon)] - \mathbb{E}^{x_0}[A(X_n)] \right| \pi^\epsilon(dx_0)$$

$$\leq R_1 \pi^\epsilon(|B|V)\epsilon \leq R_1 (\pi^\epsilon(B^2))^{\frac{1}{2}} (\pi^\epsilon(V^2))^{\frac{1}{2}} \epsilon,$$

for some constants $R_1 \in (0, +\infty)$ and $D \in (1, +\infty)$ (independent of A and B). Here we have applied Proposition 2.1 to $\left| \mathbb{E}^{x_0}[A(X_n^\epsilon)] - \mathbb{E}^{x_0}[A(X_n)] \right|$ based on Lemma 3.1.

Meanwhile for the term I_2, we have

$$I_2 = \int B(x_0) \int A(x) P_{n\delta}(x_0, dx)(\pi^\epsilon - \pi)(dx_0) = \int B(x_0) E^{x_0}[A(X_n)](\pi^\epsilon - \pi)(dx_0) = \pi^\epsilon(f_n) - \pi(f_n),$$

where $f_n(x) = B(x)\mathbb{E}^x[A(X_n)]$. Here, f_n, in general, is not a function in $\mathscr{G}_\ell$, and we cannot apply the existing one-point statistics error bound derived from the perturbation theory, e.g., Proposition 2.1. As a remedy, we consider the long-time linear response theory reviewed in Section 2.4.

Before applying Theorem 2.2 to f_n, we need to show that $f_n \in C^1_{G,H}$. Let

$$A_n(x) := \mathbb{E}^x[A(X_n)] - \pi(A).$$

Notice $A \in C^1_{V,VH/G} \subset C^1_{G,H}$ (since $G \geq V$), and, by the spectral gap assumption (Assumption 2.6), we have

$$\|A_n\|_{1;G,H} = \left\| \mathscr{P}^0_{n\delta} A - \pi(A) \right\|_{1;G,H} \leq \lambda^n \|A - \pi(A)\|_{1;G,H}, \quad \lambda \in (0, 1). \tag{75}$$

With this bound, we turn to f_n,

$$\|f_n\|_{1;G,H} = \|BA_n + B\pi(A)\|_{1;G,H} \leq \|BA_n\|_{1;G,H} + \pi(A)\|B\|_{1;G,H},$$

where $\|B\|_{1;G,H} < \infty$ since $B \in C^1_{G/V,H/V} \subset C^1_{G,H}$. Therefore, to show $f_n \in C^1_{G,H}$ it is enough to control the norm $\|BA_n\|_{1;G,H}$. By the definition (13), we have

$$\|BA_n\|_{1;G,H} = \sup_{x\in\mathbb{R}^d} \left\{ \frac{|BA_n|}{G} + \frac{\|\nabla BA_n + B\nabla A_n\|}{H} \right\}.$$

Notice that,

$$\frac{|BA_n|}{G} + \frac{\|\nabla BA_n + B\nabla A_n\|}{H} \leq |B| \left(\frac{|A_n|}{G} + \frac{\|\nabla A_n\|}{H} \right) + \frac{\|\nabla B\|}{H/G} \frac{|A_n|}{G} \leq \left(\frac{|A_n|}{G} + \frac{\|\nabla A_n\|}{H} \right) \left(|B| + \frac{\|\nabla B\|}{H/G} \right),$$

where, by Eq. (75),

$$\sup_{x\in\mathbb{R}^d} \left\{ \frac{|A_n|}{G} + \frac{\|\nabla A_n\|}{H} \right\} = \|A_n\|_{1;G,H} \leq \lambda^n \|A - \pi(A)\|_{1;G,H} = \lambda^n \sup_{x\in\mathbb{R}^d} \left\{ \frac{|A - \pi(A)|}{G} + \frac{\|\nabla A\|}{H} \right\}.$$

Thus, we obtain the following bound,

$$\|BA_n\|_{1;G,H} \leq \lambda^n \sup_{x\in\mathbb{R}^d} \left\{ \left(\frac{|A - \pi(A)|}{G} + \frac{\|\nabla A\|}{H} \right) \left(|B| + \frac{\|\nabla B\|}{H/G} \right) \right\} = \lambda^n \sup_{x\in\mathbb{R}^d} \left\{ \left(\frac{|A - \pi(A)|}{V} + \frac{\|\nabla A\|}{VH/G} \right) \left(\frac{|B|}{G/V} + \frac{\|\nabla B\|}{H/V} \right) \right\}.$$

This shows that,

$$\|f_n\|_{1;G,H} \leq \lambda^n \|A - \pi(A)\|_{1;V,VH/G} \|B\|_{1;G/V,H/V} + \pi(A)\|B\|_{1;G,H} < \infty. \tag{76}$$

With $f_n \in C^1_{G,H}$, by invoking Theorem 2.2, we have

$$|I_2| = |\pi^\epsilon(f_n) - \pi(f_n)| \leq \left| \frac{d}{d\epsilon} \pi^\epsilon(f_n) \right|_{\epsilon=0} \epsilon + O(\epsilon^2),$$

where the ϵ-derivative is well-defined and satisfies Eq. (14) with $f = f_n$ and $t = \delta$.

Given $\pi(A) = 0$, Eq. (76) reduces to

$$\|f_n\|_{1;G,H} \leq \lambda^n \|A\|_{1;V,VH/G} \|B\|_{1;G/V,H/V},$$

and the ϵ-derivative in (30), by Eq. (14), satisfies

$$\frac{d}{d\epsilon} \pi^\epsilon(f_n) \Big|_{\epsilon=0} = \mathbb{E}_\pi \left[\partial \mathscr{P}^0_\delta (I - \mathscr{P}^0_\delta)^{-1} (f_n - \pi(f_n)) \right],$$

where

$$\|f_n - \pi(f_n)\|_{1;G,H} \leq \|f_n\|_{1;G,H} + |\pi(f_n)| \leq \lambda^n \|A\|_{1;V,VH/G} \|B\|_{1;G/V,H/V} + |\pi(f_n)|, \quad \forall n \geq 0.$$

To bound $\pi(f_n)$, since $A \in \mathscr{G}_\ell \subset \mathscr{G}$, we apply Theorem 2.1 to A,

$$|f_n(x)| = |B(x)| \left| \mathbb{E}^x[A(X_n)] \right| = |B(x)| \left| \mathbb{E}^x[A(X_n)] - \pi(A) \right| \leq R_2 \rho^n |B(x)| V(x) \leq R_2 \rho^n \|B\|_{1;G/V,H/V} G(x), \quad \forall x \in \mathbb{R}^d,$$

for some constants $R_2 \in (0, \infty)$ and $\rho = \rho(\delta) \in (0, 1)$ independent of A. Here, $|B(x)|V(x) \leq \|B\|_{1;G/V,H/V} G(x)$ by the definition of the norm $\|\cdot\|_{G/V,H/V}$. Thus, $|\pi(f_n)| \leq R_2 \|B\|_{1;G/V,H/V} \rho^n \pi(G)$ and

$$\|f_n - \pi(f_n)\|_{1;G,H} \leq \|B\|_{1;G/V,H/V} \left(\lambda^n \|A\|_{1;V,VH/G} + R_2 \rho^n \pi(G) \right), \tag{77}$$

where $\pi(G) \leq \pi(U) < \infty$ by Assumptions 2.7–2.8.

Let $K_\pi := \left\{ f \in C^1_{G,H} \mid \pi(f) = 0 \right\}$. Recall that $(I - \mathscr{P}^0_\delta)^{-1}$ defines a bounded linear map from K_π to itself (see the discussion after Assumption 2.6). In particular, for any $f \in K_\pi$, we have,

$$(I - \mathscr{P}^0_\delta)^{-1} f(x) = \sum_{n=0}^\infty \left(\mathscr{P}^0_\delta \right)^n f(x) = \sum_{n=0}^\infty \mathbb{E}^x[f(X_n)]$$

(the summation converges due to the spectral gap assumption), which implies that,

$$\left\| (I - \mathscr{P}_\delta^0)^{-1} f \right\|_{1;G,H} = \left\| \sum_{n=0}^\infty \left(\mathscr{P}_\delta^0 \right)^n f(x) \right\|_{1;G,H} \leq \sum_{n=0}^\infty \lambda^n \| f \|_{1;G,H} = \frac{1}{1-\lambda} \| f \|_{1;G,H}, \quad \forall f \in K_\pi.$$

Together with Assumption 2.7 and Eq. (77), we have

$$\left\| \partial \mathscr{P}_\delta^0 (I - \mathscr{P}_\delta^0)^{-1} \left(f_n - \pi(f_n) \right) \right\|_U \leq \frac{R_3}{1-\lambda} \| f_n - \pi(f_n) \|_{1;G,H} \leq \frac{R_3}{1-\lambda} \| B \|_{1;G/V,H/V} \left(\lambda^n \| A \|_{1;V,VH/G} + R_2 \rho^n \pi(G) \right),$$

for a constant $R_3 \in (0, +\infty)$ independent of A and B, which leads to the desirable error bound in Eq. (31).

References

[1] Peter E. Kloeden, Eckhard Platen, Numerical Solution of Stochastic Differential Equations, vol. 23, Springer Science & Business Media, 2013.
[2] Bernt Oksendal, Stochastic Differential Equations: An Introduction with Applications, Springer Science & Business Media, 2013.
[3] Grigorios A. Pavliotis, Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations, vol. 60, Springer, 2014.
[4] Jan Nygaard Nielsen, Henrik Madsen, Peter C. Young, Parameter estimation in stochastic differential equations: an overview, Annu. Rev. Control 24 (2000) 83–94.
[5] Daniel Peavoy, Christian L.E. Franzke, Gareth O. Roberts, Systematic physics constrained parameter estimation of stochastic differential equations, Comput. Statist. Data Anal. 83 (2015) 182–199.
[6] Isambi S. Mbalawata, Simo Särkkä, Heikki Haario, Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering, Comput. Statist. 28 (3) (2013) 1195–1223.
[7] Xi Chen, Ilya Timofeyev, Non-parametric estimation of stochastic differential equations from stationary time-series, 2020, ArXiv E-Prints, arXiv–2007.
[8] Richard Nickl, Kolyan Ray, et al., Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions, Ann. Statist. 48 (3) (2020) 1383–1408.
[9] Constantino A. Garcia, Abraham Otero, Paulo Felix, Jesus Presedo, David G. Marquez, Nonparametric estimation of stochastic differential equations with sparse Gaussian processes, Phys. Rev. E 96 (2) (2017) 022104.
[10] Yalda Rajabzadeh, Amir Hossein Rezaie, Hamidreza Amindavar, A robust nonparametric framework for reconstruction of stochastic differential equation models, Physica A 450 (2016) 294–304.
[11] David Lamouroux, Klaus Lehnertz, Kernel-based regression of drift and diffusion coefficients of stochastic processes, Phys. Lett. A 373 (39) (2009) 3507–3512.
[12] Stefan Chmiela, Huziel E. Sauceda, Klaus-Robert Müller, Alexandre Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, Nature Commun. 9 (1) (2018) 1–10.
[13] Stefan Chmiela, Huziel E. Sauceda, Alexandre Tkatchenko, Klaus-Robert Müller, Accurate molecular dynamics enabled by efficient physically constrained machine learning approaches, in: Machine Learning Meets Quantum Physics, Springer, 2020, pp. 129–154.
[14] Shixiao W. Jiang, John Harlim, Modeling of missing dynamical systems: Deriving parametric models using a nonparametric framework, Res. Math. Sci. 7 (3) (2020) 1–25.
[15] Tyrus Berry, Dimitrios Giannakis, John Harlim, Nonparametric forecasting of low-dimensional dynamical systems, Phys. Rev. E 91 (3) (2015) 032915.
[16] Tyrus Berry, Dimitrios Giannakis, John Harlim, Bridging data science and dynamical systems theory, Notices Amer. Math. Soc. 67 (9) (2020) 1336–1349.
[17] Faheem Gilani, Dimitrios Giannakis, John Harlim, Kernel-based prediction of non-Markovian time series, Physica D 418 (2021) 132829.
[18] Debasish Koner, Markus Meuwly, Permutationally invariant, reproducing kernel-based potential energy surfaces for polyatomic molecules: From formaldehyde to acetone, J. Chem. Theory Comput. 16 (9) (2020) 5474–5484.
[19] John Harlim, Shixiao W. Jiang, Senwei Liang, Haizhao Yang, Machine learning for prediction with missing dynamics, J. Comput. Phys. (2020) 109922, http://dx.doi.org/10.1016/j.jcp.2020.109922.
[20] Pantelis R. Vlachas, Wonmin Byeon, Zhong Y. Wan, Themistoklis P. Sapsis, Petros Koumoutsakos, Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 474 (2213) (2018) 20170844.
[21] Chao Ma, Jianchun Wang, Weinan E., Model reduction with memory and the machine learning of dynamical systems, Commun. Comput. Phys. 25 (4) (2019) 947–962.
[22] Shaowu Pan, Karthik Duraisamy, Data-driven discovery of closure models, SIAM J. Appl. Dyn. Syst. 17 (4) (2018) 2381–2413.
[23] Barbara Hammer, On the approximation capability of recurrent neural networks, Neurocomputing 31 (1) (2000) 107–123.
[24] Anton Maximilian Schäfer, Hans Georg Zimmermann, Recurrent neural networks are universal approximators, in: International Conference on Artificial Neural Networks, Springer, 2006, pp. 632–640.
[25] Herbert Jaeger, The "echo state" approach to analysing and training recurrent neural networks-with an erratum note, Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148 (34) (2001) 13.
[26] Ali Rahimi, Benjamin Recht, Uniform approximation of functions with random bases, in: 2008 46th Annual Allerton Conference on Communication, Control, and Computing, IEEE, 2008, pp. 555–561.
[27] Herbert Jaeger, Harald Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science 304 (5667) (2004) 78–80.
[28] Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, Edward Ott, Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Phys. Rev. Lett. 120 (2) (2018) 024102.
[29] Jaideep Pathak, Zhixin Lu, Brian R. Hunt, Michelle Girvan, Edward Ott, Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data, Chaos 27 (12) (2017) 121102.
[30] Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, Approximation bounds for random neural networks and reservoir systems, 2020, arXiv preprint arXiv:2002.05933.
[31] Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, Risk bounds for reservoir computing, J. Mach. Learn. Res. 21 (240) (2020) 1–61.
[32] Lyudmila Grigoryeva, Juan-Pablo Ortega, Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems, J. Mach. Learn. Res. 19 (2018) 1–40.
[33] D. Crommelin, E. Vanden-Eijnden, Subgrid-scale parameterization with conditional Markov chains, J. Atmos. Sci. 65 (2008) 2661–2675.
[34] A.J. Majda, J. Harlim, Physics constrained nonlinear regression models for time series, Nonlinearity 26 (2013) 201–217.
[35] Alexandre J. Chorin, Fei Lu, Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics, Proc. Natl. Acad. Sci. 112 (32) (2015) 9804–9809.
[36] J. Harlim, X. Li, Parametric reduced models for the nonlinear Schrödinger equation, Phys. Rev. E 91 (2015) 053306.
[37] Daniel Rudolf, Nikolaus Schweizer, Perturbation theory for Markov chains via wasserstein distance, Bernoulli 24 (4A) (2018) 2610–2639.
[38] Tony Shardlow, Andrew M. Stuart, A perturbation theory for ergodic Markov chains and application to numerical approximations, SIAM J. Numer. Anal. 37 (4) (2000) 1120–1137.
[39] Jonathan C. Mattingly, Andrew M. Stuart, Desmond J. Higham, Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise, Stochastic Process. Appl. 101 (2) (2002) 185–232.
[40] Martin Hairer, Andrew J. Majda, A simple framework to justify linear response theory, Nonlinearity 23 (4) (2010) 909.
[41] Lorenzo Rosasco, Mikhail Belkin, Ernesto De Vito, On learning with integral operators, J. Mach. Learn. Res. 11 (2) (2010).
[42] Cheng Wang, Ding-Xuan Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, J. Complexity 27 (1) (2011) 55–67.
[43] Cheng Wang, Zheng Chu Guo, ERM learning with unbounded sampling, Acta Math. Sin. (Engl. Ser.) 28 (1) (2012) 97–104.

[44] Felipe Cucker, Steve Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (1) (2002) 1–49.
[45] Vladimir I. Bogachev, Nicolai V. Krylov, Michael Röckner, Stanislav V. Shaposhnikov, Fokker-Planck-Kolmogorov Equations, vol. 207, American Mathematical Soc., 2015.
[46] H. Lamba, Jonathan C. Mattingly, Andrew M. Stuart, An adaptive Euler-maruyama scheme for SDEs: Convergence and stability, IMA J. Numer. Anal. 27 (3) (2007) 479–506.
[47] Joel A. Tropp, User-friendly tail bounds for sums of random matrices, Found. Comput. Math. 12 (4) (2012) 389–434.
[48] Jonathan Huggins, James Zou, Quantifying the accuracy of approximate diffusions and Markov chains, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 382–391.
[49] John Harlim, Xiantao Li, He Zhang, A parameter estimation method using linear response statistics, J. Stat. Phys. 168 (1) (2017) 146–170.
[50] Cecil E. Leith, Climate response and fluctuation dissipation, J. Atmos. Sci. 32 (10) (1975) 2022–2026.
[51] Andrew Majda, Rafail V. Abramov, Marcus J. Grote, Information Theory and Stochastics for Multiscale Nonlinear Systems, vol. 25, American Mathematical Soc., 2005.
[52] He Zhang, John Harlim, Xiantao Li, Linear response based parameter estimation in the presence of model error, J. Comput. Phys. 430 (2021) 110112.
[53] He Zhang, John Harlim, Xiantao Li, Estimating linear response statistics using orthogonal polynomials: An RKHS formulation, Found. Data Sci. 2 (4) (2020) 443.
[54] He Zhang, Xiantao Li, John Harlim, A parameter estimation method using linear response statistics: Numerical scheme, Chaos 29 (3) (2019) 033101.
[55] Wen Huang, Min Ji, Zhenxin Liu, Yingfei Yi, Steady states of fokker-Planck equations: III. Degenerate diffusion, J. Dynam. Differential Equations 28 (1) (2016) 127–141.
[56] Ingo Steinwart, Andreas Christmann, Support Vector Machines, Springer Science & Business Media, 2008.
[57] Hongwei Sun, Mercer theorem for RKHS on noncompact sets, J. Complexity 21 (3) (2005) 337–349.
[58] Andrea Caponnetto, Ernesto De Vito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. 7 (3) (2007) 331–368.
[59] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, Dejan Slepčev, Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–beltrami operator, Found. Comput. Math. 20 (4) (2020) 827–887.
[60] Tyrus Berry, John Harlim, Variable bandwidth diffusion kernels, Appl. Comput. Harmon. Anal. 40 (1) (2016) 68–96.
[61] Boumediene Hamzi, Houman Owhadi, Learning dynamical systems from data: A simple cross-validation perspective, part I: Parametric kernel flows, Physica D (ISSN: 0167-2789) 421 (2021) 132817.
[62] Romeo Alexander, Dimitrios Giannakis, Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques, Physica D 409 (2020) 132520.
[63] Ulrike Von Luxburg, Mikhail Belkin, Olivier Bousquet, Consistency of spectral clustering, Ann. Statist. (2008) 555–586.
[64] Martin J. Wainwright, High-Dimensional Statistics: A Non-Asymptotic Viewpoint, vol. 48, Cambridge University Press, 2019.
[65] Bharath K. Sriperumbudur, Kenji Fukumizu, Gert R.G. Lanckriet, Universality, characteristic kernels and RKHS embedding of measures, J. Mach. Learn. Res. 12 (7) (2011).
[66] Ding-Xuan Zhou, The covering number in learning theory, J. Complexity 18 (3) (2002) 739–767.
[67] Cem Anil, James Lucas, Roger Grosse, Sorting out Lipschitz function approximation, in: International Conference on Machine Learning, PMLR, 2019, pp. 291–301.
[68] Lai-Sang Young, What are SRB measures, and which dynamical systems have them? J. Stat. Phys. 108 (5) (2002) 733–754.
[69] Vivien Baladi, Linear response, or else, in: Proceedings of ICM, Seoul, vol. III, 2014, pp. 525–545.
[70] Sever Silvestru Dragomir, Melbourne City, Some Grönwall type inequalities and applications, 2002, URL: http://rgmia.vu.edu.au/SSDragomirWeb.html.
[71] Gene H. Golub, Charles F. Van Loan, Matrix Computations, vol. 3, JHU Press, 2013.
[72] Zlatko Drmač, Matjaž Omladič, Krešimir Veselić, On the perturbation of the cholesky factorization, SIAM J. Matrix Anal. Appl. 15 (4) (1994) 1319–1332.